

# Protein interaction networks and biology: towards the connection

A Annibale<sup>†</sup>, ACC Coolen<sup>†‡§</sup>, N Planell-Morell<sup>†</sup>

<sup>†</sup> Department of Mathematics, King's College London, The Strand, London WC2R 2LS, UK

<sup>‡</sup> Institute for Mathematical and Molecular Biomedicine, King's College London, Hodgkin Building, London SE1 1UL, UK

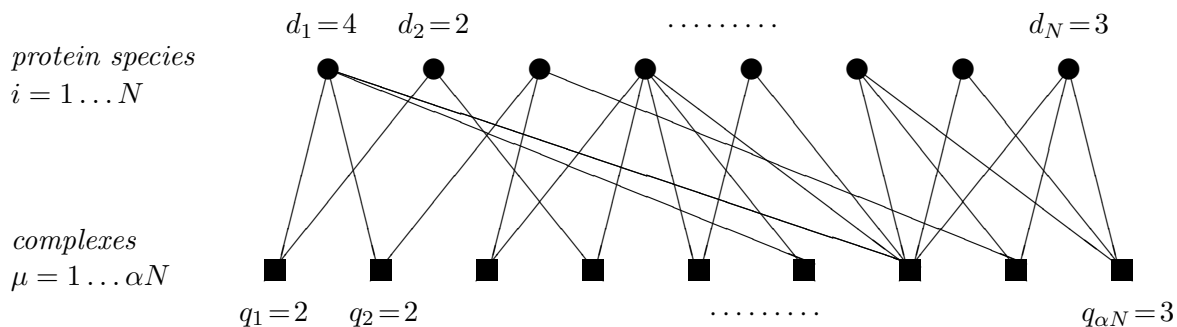
<sup>§</sup> London Institute for Mathematical Sciences, 22 South Audley St, London W1K 2NY, UK

**Abstract.** Protein interaction networks (PIN) are popular means to visualize the proteome. However, PIN datasets are known to be noisy, incomplete and biased by the experimental protocols used to detect protein interactions. This paper aims at understanding the connection between true protein interactions and the protein interaction datasets that have been obtained using the most popular experimental techniques, i.e. mass spectrometry (MS) and yeast two-hybrid (Y2H). We show that the most natural adjacency matrix of protein interaction networks has a separable form, and this induces precise relations between moments of the degree distribution and the number of short loops. These relations provide powerful tools to test the reliability of datasets and hint at the underlying biological mechanism with which proteins and complexes recruit each other.

## 1. Introduction

A protein interaction network (PIN) is a graph where nodes  $i = 1 \dots N$  represent proteins and links represent their interactions. This graph is encoded in an adjacency matrix  $\mathbf{a} = \{a_{ij}\}$ , whose entries denote whether there is a link between proteins  $i$  and  $j$  ( $a_{ij}=1$ ) or not ( $a_{ij}=0$ ). However, there is ambiguity in its definition, arising from the non-binarity of the underlying biochemistry. For example, three proteins may form a complex, but may not interact in pairs. Assigning binary values to intrinsically non-binary interactions requires further prescriptions, which vary across experimental protocols and lead in practice to different graphs. Moreover, different experiments measure protein interactions in different ways, which causes further biases [1, 2, 3]. For quantitative studies of the effects of sampling biases on networks see e.g. [4, 6, 7, 5, 8, 9, 10].

In this paper we seek to establish the connection between true biological protein interactions and protein interaction datasets produced by the most popular experimental techniques, mass spectrometry (MS) and yeast two-hybrid (Y2H). We argue that the most natural network matrix representation of the proteome has a separable form, which induces precise relations between the degree distribution and the density of short loops. These relations provide simple tests to assess the reliability and quality of different data sets, and provide hints on the underlying (evolutionary) mechanisms with which proteins and complexes recruit each other. Our study also provides a theoretical framework to discriminate between ‘party’ and ‘date’ hubs in protein interaction networks, see e.g. [17] and references therein, and addresses several intriguing



**Figure 1.** Bipartite graph (or ‘factor graph’) representation of protein interactions. The protein species  $i = 1 \dots N$  are drawn as circles, and their complexes  $\mu = 1 \dots \alpha N$  as squares. We write the degree of protein  $i$  as  $d_i$  (the number of complexes it participates in), and the degree of complex  $\mu$  as  $q_\mu$  (the number of protein species it contains). The bipartite graph gives more detailed information than the conventional PIN with protein nodes and pairwise links only. For instance, one distinguishes easily between different types of ‘hub’ proteins: ‘date hub’ proteins connect to many degree-2 complexes, whereas ‘party hub’ proteins connect to a high degree complex.

questions concerning the universality of protein and complex statistics across species. For example, given  $N$  protein species in a cell, what is the number of complexes they typically form, i.e. to what extent is the ratio complexes/proteins conserved across different species? Is the distribution of complex sizes peaked around ‘typical’ values, or does it have long tails? How is this mirrored in the protein promiscuities, i.e. the propensities of proteins to participate in multiple complexes? Does the power law behaviour of the degree distribution of protein interaction networks perhaps result from tails in the distribution of complex sizes and protein promiscuities?

We tackle the above questions using an approach that is entirely based on statistical properties of graph ensembles. In section 2 we first define our models. Sections 3, 4 and 5 are devoted to the derivation of properties of distinct separable graph ensembles which mimic protein interaction networks, each reflecting different possible mechanisms for complex genesis. In section 6 we test these properties in synthetically generated graphs, and in section 7 we do the same for protein interaction networks measured by MS and Y2H experiments. We end our paper with a summary of our conclusions, and suggest pathways for further research.

## 2. Definitions and basic properties

### 2.1. The bipartite graph representation of the proteome

Proteins are large and complicated heteropolymers, which can bind in specific combinations to form stable molecular complexes. We consider a set of  $N$  protein species, labelled by  $i = 1 \dots N$ . We assume that the number of stable complexes  $p$  scales as  $p = \alpha N$  where  $\alpha > 0$ , and we label the complexes by  $\mu = 1 \dots \alpha N$ . We can represent this system as a bi-partite graph [11], see Figure 1, with two sets of nodes. The set  $\nu_p$  represents proteins (drawn as circles), the set  $\nu_c$  represents complexes (drawn as squares), and a link between protein  $i \in \nu_p$  and complex  $\mu \in \nu_c$  is drawn if protein  $i$  participates in complex  $\mu$ . This graph is defined by the  $N \times \alpha N$  connectivity matrix

$\xi = \{\xi_i^\mu\}$ , where  $\xi_i^\mu = 1$  if there is a link between  $i$  and  $\mu$ , and  $\xi_i^\mu = 0$  otherwise. For simplicity we do not allow for complexes with more than one occurrence of any given protein species.

In the bipartite graph one has two types of node degrees: the degree  $d_i(\xi) = \sum_\mu \xi_i^\mu$  (or ‘promiscuity’) of each protein  $i$  gives the number of different complexes in which it is involved, and the degree  $q_\mu(\xi) = \sum_i \xi_i^\mu$  (or ‘size’) of each complex  $\mu$  gives the number of protein species of which it is formed. We define the distribution of promiscuities in graph  $\xi$  as  $p(d|\xi) = N^{-1} \sum_i \delta_{d,d_i(\xi)}$ , with the average promiscuity  $\langle d(\xi) \rangle = \sum_d dp(d|\xi)$ , and the distribution of complex sizes as  $p(q|\xi) = (\alpha N)^{-1} \sum_{\mu=1}^{\alpha N} \delta_{q,q_\mu(\xi)}$ , with the average complex size  $\langle q(\xi) \rangle = \sum_q qp(q|\xi)$ . Since the number of links is conserved, we always have  $\langle d(\xi) \rangle = \alpha \langle q(\xi) \rangle$  for any bipartite graph  $\xi$ .

## 2.2. Link distribution in the bipartite graph

Since we generally do not know the microscopic bipartite graph  $\xi$ , we will regard it as a quenched random object. Several natural choices can be proposed for its distribution  $p(\xi)$ . If we assume that complexes recruit proteins, independently and with the same likelihood, we are led to

$$p_A(\xi) = \prod_{i\mu} \left[ \frac{q_\mu}{N} \delta_{\xi_i^\mu,1} + \left( 1 - \frac{q_\mu}{N} \right) \delta_{\xi_i^\mu,0} \right] \quad (1)$$

with  $\delta_{xy} = 1$  for  $x = y$  and 0 otherwise, and where the  $\{q_\mu\}$  are distributed according to  $P(q) = (\alpha N)^{-1} \sum_\mu \delta_{q,q_\mu}$ . For graphs  $\xi$  drawn from the ensemble (1) and  $N \rightarrow \infty$ , each complex size  $q_\mu(\xi)$  is a Poissonian random variable with average  $q_\mu$ , and all protein promiscuities  $d_i(\xi)$  are Poissonian variables with average  $\langle d \rangle = \alpha \langle q \rangle$ , since

$$\begin{aligned} p(d) &= \lim_{N \rightarrow \infty} \langle \delta_{d, \sum_\mu \xi_i^\mu} \rangle = \lim_{N \rightarrow \infty} \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega d} \langle e^{-i\omega \sum_\mu \xi_i^\mu} \rangle \\ &= \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega d + \alpha \langle q \rangle (e^{-i\omega} - 1)} = e^{-\alpha \langle q \rangle} (\alpha \langle q \rangle)^d / d! \end{aligned} \quad (2)$$

In the scenario (1) complexes have sizes that are determined e.g. by their functions, and this controls the promiscuities of the recruited proteins. Alternatively one could assume that the likelihood of a protein participating in a complex is driven by its promiscuity, leading to the ‘dual’ ensemble

$$p_B(\xi) = \prod_{i\mu} \left[ \frac{d_i}{\alpha N} \delta_{\xi_i^\mu,1} + \left( 1 - \frac{d_i}{\alpha N} \right) \delta_{\xi_i^\mu,0} \right] \quad (3)$$

where the  $\{d_i\}$  are distributed according to  $P(d) = N^{-1} \sum_i \delta_{d,d_i}$ . Here as  $N \rightarrow \infty$  the protein promiscuities  $d_i(\xi)$  are Poissonian variables with averages  $d_i$ , whereas all complex sizes  $q_\mu(\xi)$  are Poisson variables with identical average  $\langle q \rangle = \langle d \rangle / \alpha$ , since

$$\begin{aligned} p(q) &= \lim_{N \rightarrow \infty} \langle \delta_{q, \sum_i \xi_i^\mu} \rangle = \lim_{N \rightarrow \infty} \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega q} \langle e^{-i\omega \sum_i \xi_i^\mu} \rangle = \\ &= \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega q + \frac{\langle d \rangle}{\alpha} (e^{-i\omega} - 1)} = e^{-\langle d \rangle / \alpha} (\langle d \rangle / \alpha)^q / q! \end{aligned} \quad (4)$$

In this second ensemble proteins have intrinsic promiscuities, determined e.g. by the number of their binding sites, their polarization and so on, and these drive their recruitment to complexes.

A third obvious choice is the ‘mixed’ ensemble

$$p_C(\boldsymbol{\xi}) = \prod_{i\mu} \left[ \frac{d_i q_\mu}{\alpha N \langle q \rangle} \delta_{\xi_i^\mu, 1} + \left( 1 - \frac{d_i q_\mu}{\alpha N \langle q \rangle} \right) \delta_{\xi_i^\mu, 0} \right] \quad (5)$$

where all protein promiscuties and complex sizes are constrained on average, i.e.  $\langle d_i(\boldsymbol{\xi}) \rangle = d_i$  and  $\langle q_\mu(\boldsymbol{\xi}) \rangle = q_\mu$ , with  $\{d_i\}$  and  $\{q_\mu\}$  distributed according to  $P(d)$  and  $P(q)$ . Here protein binding statistics are driven both by complex functionality and protein promiscuity factors. The mixed ensemble (5) reduces to (1) for the choice  $P(d) = \delta_{d, \alpha \langle q \rangle}$ , and to (3) when  $P(q) = \delta_{q, \langle q \rangle}$ . By determining which of the above ensemble reflects better biological reality, we will thus learn about the mechanisms with which complexes and proteins recruit each other.

The above three ensembles become equivalent when  $q_\mu = \langle q \rangle \forall \mu$  and  $d_i = \alpha \langle q \rangle \forall i$ . In that case complex sizes and protein promiscuties are homogeneous, and the recruitment process between proteins and complexes is fully random. Bipartite graphs drawn from (1) were found to have modular topologies, and to accomplish parallel information processing for suitable values of the parameter  $\alpha$  [14, 12]. Their ensemble entropy has been calculated in [15]. One can show easily that if one replaces the soft constraints on the local degrees in our soft-constrained graph ensembles (1,3) by hard constraints, then one finds asymptotically the same distributions (2,4). Finally, we note that all three ensembles (1,3,5) are of the form  $p(\boldsymbol{\xi}) = \prod_{i\mu} p_{i\mu}(\xi_i^\mu)$ , so there are no correlations between the entries of  $\boldsymbol{\xi}$ . This strong assumption of our models will need to be checked a posteriori.

### 2.3. Accounting for binding sites

In all PINs each protein is reduced to a simple network node, in spite of the fact that proteins are in reality complex chains of aminoacids with several binding domains. Here we show that the ensembles introduced in the previous section can accommodate the presence of multiple binding sites when these are equally reactive. Let us first assume that each protein has  $d$  functional reactive amino-acid endgroups. When two such proteins bind, the resulting dimer has  $2d - 2$  unused reactive endgroups, a trimer has  $3d - 4$  endgroups, and a  $k$ -mer has  $kd - 2(k - 1) = (d - 2)k + 2$  endgroups. If all endgroups are equally reactive, the a priori probability that a protein  $i$  is part of a complex  $\mu$  is given by

$$p(\xi_i^\mu = 1) = \frac{d[(d - 2)q_\mu + 2]}{Z} \simeq \frac{q_\mu d}{\alpha N \langle q \rangle} \quad (6)$$

where the last approximate equality holds for  $d \gg 1$  and  $Z = \sum_\mu q_\mu d = \alpha N \langle q \rangle d$ . This corresponds to ensemble (1), with the choice  $d = \alpha \langle q \rangle$ . If proteins have different endgroups  $d_i$ ,

$$p(\xi_i^\mu = 1) \simeq \frac{d_i[(d - 2)q_\mu + 2]}{\alpha N \langle q \rangle d} \simeq \frac{d_i q_\mu}{\alpha N \langle q \rangle} \quad (7)$$

where  $d = N^{-1} \sum_i d_i$ , leading to ensemble (5). If the variability of  $q_\mu$  is small,  $q_\mu \simeq \langle q \rangle$ ,

$$p(\xi_i^\mu = 1) = \frac{d_i}{\alpha N} \quad (8)$$

and we retrieve (3). The assumption of unbiased interactions between proteins with varying individual binding affinities has been supported in [23].

### 2.4. Protein interactions as detected by experiments

Protein detection experiments seek to measure for each pair  $(i, j)$  of protein species whether they interact in any complex, and assign an undirected link between nodes  $i$  and  $j$  if they do. Hence the PIN adjacency matrix  $\mathbf{a} = \{a_{ij}\}$  resulting from such experiments can be expressed in terms of the entries of the bipartite graph  $\boldsymbol{\xi}$  in Figure 1 via

$$a_{ij} = \theta\left(\sum_{\mu=1}^{\alpha N} \xi_i^\mu \xi_j^\mu\right) \quad \forall i \neq j \quad (9)$$

and  $a_{ii} = 0 \quad \forall i$ , with the convention  $\theta(0) = 0$  for the step function, defined by  $\theta(x > 0) = 1$  and  $\theta(x < 0) = 0$ . The aim of this paper hence translates into studying the properties of the following ensemble of nondirected random graphs, in which the  $\{\xi_i^\mu\}$  are drawn from either of the ensembles (1,3,5):

$$p(\mathbf{a}) = \left\langle \left[ \prod_{i < j} \delta_{a_{ij}, \theta(\sum_{\mu \leq \alpha N} \xi_i^\mu \xi_j^\mu)} \right] \left[ \prod_i \delta_{a_{ii}, 0} \right] \right\rangle_{\boldsymbol{\xi}} \quad (10)$$

Some properties of (1,3) will turn out not to depend on the choices made for the distributions of complex sizes and protein promiscuities, and this leads to powerful benchmarks against which to test available PIN datasets. A key feature we exploit in our analysis is that averages over (10) can often be replaced by averages over the following related ensemble of *weighted* graphs

$$p(\mathbf{c}) = \left\langle \left[ \prod_{i < j} \delta_{c_{ij}, \sum_{\mu \leq \alpha N} \xi_i^\mu \xi_j^\mu} \right] \left[ \prod_i \delta_{c_{ii}, 0} \right] \right\rangle_{\boldsymbol{\xi}} \quad (11)$$

Here an entry  $c_{ij} = \sum_{\mu \leq \alpha N} \xi_i^\mu \xi_j^\mu \in \mathbb{N}$  represents the *number* of complexes in which proteins  $i$  and  $j$  participate simultaneously. For finite  $q_\mu, d_i$  and  $\alpha$ , one finds that in large networks generated via (1,3,5) the probability of seeing  $c_{ij} > 1$  is of order  $\mathcal{O}(N^{-2})$ , and the values of many macroscopic observables in the  $\mathbf{a}$  and  $\mathbf{c}$  ensembles will, to leading order in  $N$ , be identical.

## 3. Network properties generated by the $q$ -ensemble

In this section we study the statistical properties of the ensembles (11) and (10) upon generating the bipartite protein interaction graph  $\boldsymbol{\xi}$  from ensemble (1), where complexes recruit proteins.

### 3.1. Link probabilities

For the graphs  $\mathbf{c}$  of (11) we find the following expectation values of individual bonds

$$\langle c_{ij} \rangle = \sum_{\mu=1}^{\alpha N} \langle \xi_i^\mu \xi_j^\mu \rangle_{\boldsymbol{\xi}} = \sum_{\mu=1}^{\alpha N} \left( \frac{q_\mu}{N} \right)^2 = \frac{\alpha}{N} \langle q^2 \rangle \quad (12)$$

where the brackets on the right-hand side denote averaging over the complex size distribution  $P(q)$ . The likelihood of an individual bond is (see Appendix A)

$$\begin{aligned} p(c_{ij}) &= \langle \delta_{c_{ij}, \sum_{\mu \leq \alpha N} \xi_i^\mu \xi_j^\mu} \rangle_{\boldsymbol{\xi}} \\ &= \delta_{c_{ij}, 0} + \frac{\alpha \langle q^2 \rangle}{N} (\delta_{c_{ij}, 1} - \delta_{c_{ij}, 0}) + \left( \frac{\alpha^2 \langle q^2 \rangle^2}{2N^2} - \frac{1}{2} \frac{\alpha \langle q^4 \rangle}{N^3} \right) (\delta_{c_{ij}, 2} - 2\delta_{c_{ij}, 1} + \delta_{c_{ij}, 0}) \end{aligned}$$

$$+ \frac{\alpha^3 \langle q^2 \rangle^3}{6N^3} (\delta_{c_{ij},3} - 3\delta_{c_{ij},2} + 3\delta_{c_{ij},1} - \delta_{c_{ij},0}) + \mathcal{O}(N^{-4}) \quad (13)$$

so we find for the first few probabilities:

$$p(0) = 1 - \frac{\alpha \langle q^2 \rangle}{N} + \frac{\alpha^2 \langle q^2 \rangle^2}{2N^2} - \frac{\alpha \langle q^4 \rangle}{2N^3} - \frac{\alpha^3 \langle q^2 \rangle^3}{6N^3} + \mathcal{O}(N^{-4}) \quad (14)$$

$$p(1) = \frac{\alpha \langle q^2 \rangle}{N} - \frac{\alpha^2 \langle q^2 \rangle^2}{N^2} + \frac{\alpha \langle q^4 \rangle}{N^3} + \frac{\alpha^3 \langle q^2 \rangle^3}{2N^3} + \mathcal{O}(N^{-4}) \quad (15)$$

and hence

$$\sum_{\ell > 1} p(\ell) = 1 - p(0) - p(1) = \mathcal{O}(N^{-2}), \quad \sum_{\ell > 1} \ell p(\ell) = \langle c_{ij} \rangle - p(1) = \mathcal{O}(N^{-2}) \quad (16)$$

The probability to have  $c_{ij} \neq 0$  is of order  $\mathcal{O}(N^{-1})$ , so the graphs generated by (11) are finitely connected. Moreover, although the graphs  $\mathbf{c}$  are in principle weighted, for large  $N$  the number of links per node that are not in  $\{0, 1\}$  will be vanishingly small.

### 3.2. Densities of short loops

We now turn to the calculation of expectation values for different observables in ensemble (11). First, we calculate the average number of ordered and oriented loops of length 3 per node, which are (see Appendix A):

$$m_3 = \left\langle \frac{1}{N} \sum_{ijk} c_{ij} c_{jk} c_{ki} \right\rangle_{\xi} = \frac{1}{N} \sum_{\mu \nu \rho=1}^{\alpha N} \sum_{i \neq j \neq k} \left\langle \xi_i^\mu \xi_j^\mu \xi_j^\nu \xi_k^\nu \xi_k^\rho \xi_i^\rho \right\rangle_{\xi} \quad (17)$$

$$= \alpha \langle q^3 \rangle + \mathcal{O}(N^{-1}) \quad (18)$$

Calculating the density of loops  $m_L$  for lengths  $L > 3$  can be simplified by returning to the bipartite graph  $\xi$ . We define a star  $S_n$  to be a simple  $(n+1)$ -node tree in  $\xi$ , of which the central node belongs to  $\nu_c$  (the complexes), and the  $n$  leaves belong to  $\nu_p$  (the proteins). Thus  $S_2$  stars represent protein dimers,  $S_3$  stars represent protein trimers, and so on. Each link in  $\mathbf{c}$  corresponds to at least one  $S_2$  star in the bipartite graph (which, in turn, can be a subset of any  $S_n$  star with  $n > 2$ ). Therefore, the total number of  $S_2$  stars in the bipartite graph,

$$\sum_{\mu} \sum_{i \neq j} \langle \xi_i^\mu \xi_j^\mu \rangle = \sum_{\mu} \sum_{i \neq j} \langle \xi_i^\mu \rangle \langle \xi_j^\mu \rangle = \sum_{i \neq j} \sum_{\mu} \frac{q_\mu^2}{N^2} = \alpha(N-1) \langle q^2 \rangle \quad (19)$$

has to equate in leading order the total number of links  $N \langle k \rangle$  in graph  $\mathbf{c}$ , yielding

$$\langle q^2 \rangle = \frac{\langle k \rangle}{\alpha} + \mathcal{O}(N^{-1}) \quad (20)$$

which is indeed in agreement with the result of the direct calculation  $\langle k \rangle = N^{-1} \sum_{ij} \langle c_{ij} \rangle$ , using (12). Similarly we can obtain the number of loops of length 3, calculated earlier, by realising that these loops arise when we have in the bipartite graph either a star  $S_3$  (which can be a subset of any  $S_n$  with  $n > 3$ ) or a combination of three  $S_2$  stars, where every leaf is shared by two stars. The contribution of the number of  $S_3$  stars per node to the number of loops of length 3 is

$$\begin{aligned} \frac{1}{N} \sum_{\mu} \sum_{i \neq j \neq k (\neq i)} \langle \xi_i^\mu \xi_j^\mu \xi_k^\mu \rangle &= \frac{1}{N} \sum_{\mu} \sum_{i \neq j \neq k (\neq i)} \langle \xi_i^\mu \rangle \langle \xi_j^\mu \rangle \langle \xi_k^\mu \rangle \\ &= \frac{1}{N} \sum_{\mu} \sum_{i \neq j \neq k (\neq i)} \frac{q_\mu^3}{N^3} = \alpha \langle q^3 \rangle + \mathcal{O}(N^{-1}) \end{aligned} \quad (21)$$

The contribution of the combination of three  $S_2$  stars, where each leaf is shared by two stars, is

$$\frac{1}{N} \sum_{[\mu, \nu, \rho]} \sum_{[i, j, k]} \langle \xi_i^\mu \xi_j^\mu \xi_j^\nu \xi_k^\nu \xi_k^\rho \xi_i^\rho \rangle = \frac{1}{N} \sum_{[\mu, \nu, \rho]} \sum_{[i, j, k]} \frac{q_\mu^2 q_\nu^2 q_\rho^2}{N^6} = \frac{1}{N} \alpha^3 \langle q^2 \rangle^3 + \mathcal{O}(N^{-1}) \quad (22)$$

with the square brackets  $[i, j, k]$  denoting that the three indices are distinct. The expected density of length-3 loops is the sum of an  $\mathcal{O}(1)$  contribution from  $S_3$  stars, plus an  $\mathcal{O}(N^{-1})$  contribution from combinations of three  $S_2$  stars that share leaves. For large  $N$  the second contribution vanishes, and we recover  $m_3 = \alpha \langle q^3 \rangle$ . Likewise, the  $\mathcal{O}(1)$  contribution to the density of length-4 loops comes from  $S_4$  stars in the bi-partite graph, which consist of five sites (four leaves and one central node) and four links, each with probability  $\mathcal{O}(N^{-1})$ . Combinations of two  $S_3$  stars with two shared leaves, or of  $S_2$  stars, always involve a number of links at least equal to the number of nodes and therefore yield sub-leading contributions. Hence, the density of loops of length 4 is

$$m_4 = \frac{1}{N} \sum_{\mu} \sum_{[i, j, k, \ell]} \langle \xi_i^\mu \xi_j^\mu \xi_k^\mu \xi_\ell^\mu \rangle = \alpha \langle q^4 \rangle + \mathcal{O}(N^{-1}) \quad (23)$$

More generally, the average density of loops of arbitrary length  $L$  is given by

$$m_L = \alpha \langle q^L \rangle + \mathcal{O}(N^{-1}) \quad (24)$$

For large  $N$  the ratio  $\alpha$  and the distribution  $P(q)$  of complex sizes apparently determine in full the statistics of loops of arbitrary length in  $\mathbf{c}$ , if the protein interactions are described by (1).

Finally, we note that if  $m_L$  gives the number of ordered and oriented loops of length  $L$  per node, the number of unordered and unoriented closed paths of length  $L$  equals  $\bar{m}_L = m_L/6$ , since there are  $L$  possible nodes to start a closed path from, and two possible orientations.

### 3.3. The degree distribution

It follows from (20, 24) that by measuring the average degree  $\langle k \rangle$  and the densities  $m_L$  of loops of length  $L$  we can compute all the moments of the distribution of complex sizes  $P(q)$ :

$$\langle q^2 \rangle = \langle k \rangle / \alpha, \quad \forall L > 2 : \quad \langle q^L \rangle = m_L / \alpha \quad (25)$$

This would allow us to calculate  $P(q)$  in full via its generating function, provided  $\alpha$  and  $\langle q \rangle$  are known. However, counting the number of loops of arbitrary length in a graph is computationally challenging, and  $\alpha$  and  $\langle q \rangle$  are generally unknown. However, it is possible to express  $P(q)$  for large  $N$  in terms of the degree distribution  $p(k)$  of  $\mathbf{c}$ . Specifically, in Appendix B we show that

$$\lim_{N \rightarrow \infty} p(k) = \int_0^\infty dy \, P(y) \, e^{-y} y^k / k! \quad (26)$$

where

$$P(y) = e^{-\alpha \langle q \rangle} \sum_{\ell \geq 0} \frac{(\alpha \langle q \rangle)^\ell}{\ell!} \sum_{q_1 \dots q_\ell \geq 0} W(q_1) \dots W(q_\ell) \delta[y - \sum_{r \leq \ell} q_r] \quad (27)$$

and  $W(q) = qP(q)/\langle q \rangle$  is the likelihood to draw a link attached to a complex-node of degree  $q$  in the bipartite graph  $\mathbf{\xi}$ . Formula (26) is easily interpreted. The degree of node  $i$  in  $\mathbf{c}$  is given by the second neighbours of  $i$  in  $\mathbf{\xi}$ ; the number  $\ell$  of first neighbours of node  $i$  will thus be a Poissonian variable with average  $\alpha \langle q \rangle$ , and each of its  $\ell$  first neighbours will have a degree  $q_r$  drawn from

$W(q_r)$ . Clearly, any tail in the distribution  $W(q)$  will induce a tail in the distribution  $p(k)$ , with (as we will show below) the same exponent, but an amplitude that is reduced by a factor  $\alpha\langle q \rangle$ .

One can complement (26) with a reciprocal relation that gives  $P(q)$  in terms of  $p(k)$ . To achieve this we define the generating functions  $Q_1(z) = \sum_k p(k)e^{-kz}$ ,  $Q_2(z) = \int_0^\infty dy P(y)e^{-yz}$  and  $Q_3(z) = \sum_q W(q)e^{-zq}$ . We then see from expression (26) for  $p(k)$  that

$$Q_1(z) = \int_0^\infty dy P(y) e^{-y} \sum_{k \geq 0} \frac{(ye^{-z})^k}{k!} = \int_0^\infty dy P(y) e^{y[e^{-z}-1]} = Q_2(1 - e^{-z}) \quad (28)$$

$$\begin{aligned} Q_2(z) &= e^{-\alpha\langle q \rangle} \sum_{\ell \geq 0} \frac{(\alpha\langle q \rangle)^\ell}{\ell!} \sum_{q_1 \dots q_\ell \geq 0} W(q_1) \dots W(q_\ell) e^{-z \sum_{r \leq \ell} q_r} \\ &= e^{-\alpha\langle q \rangle} \sum_{\ell \geq 0} \frac{(\alpha\langle q \rangle Q_3(z))^\ell}{\ell!} = e^{\alpha\langle q \rangle [Q_3(z) - 1]} \end{aligned} \quad (29)$$

The first identity can be rewritten as  $Q_1(-\log(1 - y)) = Q_2(y)$ . Inserting this into (29), allows us to express the desired  $Q_3(z)$  as

$$Q_3(z) = 1 + \frac{\log Q_2(z)}{\alpha\langle q \rangle} = 1 + \frac{\log Q_1(-\log(1 - z))}{\alpha\langle q \rangle} \quad (30)$$

which translates into

$$\sum_{q > 0} P(q) q e^{-zq} = \langle q \rangle + \frac{1}{\alpha} \log \sum_k p(k) (1 - z)^k \quad (31)$$

We can now extract the asymptotic form of  $P(q)$  from that of  $p(k)$ . The generating functions  $Q_1(z)$  of degree distributions that exhibit prominent tails, i.e.  $p(k) \simeq Ck^{-\mu}$  for large  $k$  with  $2 < \mu < 3$  (as observed in protein interaction networks [19, 18, 22, 20]), are for small  $z$  of the form

$$Q_1(z) = 1 - \langle k \rangle z + C\Gamma(1 - \mu)z^{\mu-1} + \dots \quad (32)$$

where  $\Gamma$  is Euler's gamma function [21]. For small  $z$  we may use  $1 - z \simeq e^{-z}$  to rewrite (30) as

$$\log Q_1(z) \simeq \alpha\langle q \rangle [Q_3(z) - 1] \quad (33)$$

Combining this with (32) then gives, for small  $z$ ,

$$-\langle k \rangle z + C\Gamma(1 - \mu)z^{\mu-1} \simeq \alpha\langle q \rangle [Q_3(z) - 1] \quad (34)$$

Hence, for small  $z$ ,  $Q_3(z)$  has the same form as  $Q_1(z)$ ,

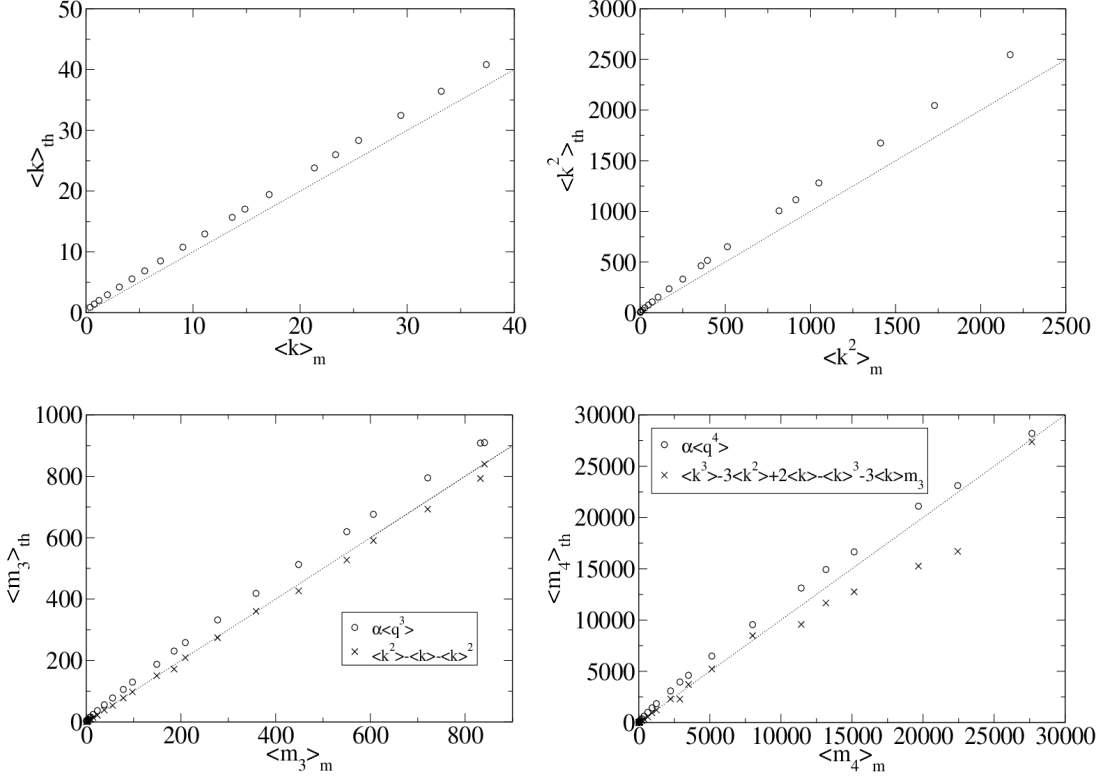
$$Q_3(z) = 1 - \frac{\langle k \rangle}{\alpha\langle q \rangle} z + \frac{C}{\alpha\langle q \rangle} \Gamma(1 - \mu) z^{\mu-1} \quad (35)$$

Therefore  $W(q)$  behaves asymptotically in the same way as  $p(k)$ , i.e.  $W(q) \simeq (C/\alpha\langle q \rangle)q^{-\mu}$ . This, in turn, gives

$$P(q) \simeq (C/\alpha) q^{-\mu-1} \quad (36)$$

The complex size distribution  $P(q)$  in (1) decays faster than the degree distribution of the associated  $\mathbf{c}$ , so fat tails in the degree distribution of protein interaction networks can emerge from less heterogeneous complex size distributions. In particular, complex size distributions with a finite second moment (but diverging higher moments) give scale-free degree distributions in  $\mathbf{c}$ . This is consistent with the intuition that, while large hubs are often observed in protein interaction





**Figure 2.** Symbols: theoretical  $\langle \dots \rangle_{\text{th}}$  versus measured  $\langle \dots \rangle_{\text{m}}$  values of observables  $\langle k \rangle$ ,  $\langle k^2 \rangle$ ,  $m_3$  and  $m_4$  in synthetically random graphs  $\mathbf{c}$  with  $N = 3000$ , defined via (1,11) for a power-law distributed complex size distribution  $P(q)$ . Theoretical values are given by formulae (37) for  $\langle k \rangle$ , (38) for  $\langle k^2 \rangle$ , (24) and (40) for  $m_3$  and (24) and (41) for  $m_4$ . Dotted lines: the diagonals (as a guides to the eye).

networks, super-complexes of the same number of proteins are unlikely to be stable. Indeed, many interactions in hubs are ‘date’ type, as opposed to ‘party’ type [17]. Our framework allows us to discriminate between different type of hub proteins, and suggests that heterogeneities in PINs may emerge from homogeneous protein ‘dating’ and moderately heterogenous protein ‘partying’.

### 3.4. Relations that are independent of $P(q)$ and $\alpha$

The first two moments of  $p(k)$  are given, to leading order in  $N$ , by (see Appendix B)

$$\langle k \rangle = \alpha \langle q^2 \rangle + \mathcal{O}(N^{-1}) \quad (37)$$

which is in agreement with (20), and

$$\langle k^2 \rangle = \alpha \langle q^2 \rangle + \alpha \langle q^3 \rangle + \alpha^2 \langle q^2 \rangle^2 \quad (38)$$

The latter is easily interpreted in terms of the underlying bipartite graph:  $\langle k^2 \rangle$  is the average density of paths of length two, so it has a contribution from  $\langle k \rangle = \alpha \langle q^2 \rangle$  due to backtracking, plus a contribution from pairs of  $S_2$  stars that share a node, whose density is

$$\frac{1}{N} \sum_{[ijk]} \sum_{\mu \neq \nu} \langle \xi_i^\mu \xi_j^\mu \xi_j^\nu \xi_k^\nu \rangle = \frac{1}{N} \sum_{[ijk]} \sum_{\mu \neq \nu} \frac{q_\mu^2}{N^2} \frac{q_\nu^2}{N^2} = \alpha^2 \langle q^2 \rangle^2, \quad (39)$$

plus a contribution from  $S_3$  stars, whose density is  $\alpha\langle q^3 \rangle$  (as shown earlier). Combining (38) with (25) gives us a relation between average and width of the degree distribution of  $\mathbf{c}$  and its density of length-3 loops. Remarkably, this relation is completely independent of  $\alpha$  and  $P(q)$ :

$$m_3 = \langle k^2 \rangle - \langle k \rangle^2 - \langle k \rangle \quad (40)$$

This identity and others, which all depend only on the separable underlying nature of the PIN and the assumption of complex-driven recruitment of proteins to complexes, can be derived more systematically from (31) by expanding both sides as power series in  $z$  and comparing the expansion coefficients. This gives a hierarchy of relations between moments of  $p(k)$  and  $P(q)$ , and hence (via (24)) between moments of  $p(k)$  and densities of loops of increasing length, that are all completely independent of  $\alpha$  and  $P(q)$ . At order  $z^2$  one recovers (40). The next order  $z^3$  leads to

$$\begin{aligned} m_4 &= \langle k^3 \rangle - 3\langle k^2 \rangle + 2\langle k \rangle + \langle k \rangle(\langle k^2 \rangle - \langle k \rangle - 2\langle k \rangle^2) \\ &= \langle k^3 \rangle - 3\langle k^2 \rangle + 2\langle k \rangle - \langle k \rangle^3 - 3\langle k \rangle m_3 \end{aligned} \quad (41)$$

To test these asymptotic identities in finite systems, we generate random graphs  $\mathbf{c}$  of size  $N = 3000$  according to (1,11), and we compared the measured values of  $m_3$  and  $m_4$  in these random graphs with the predictions of formulae (40) and (41), respectively. We show the results in Figure 2.

### 3.5. Link between $\mathbf{a}$ and $\mathbf{c}$ graph definitions

In conventional experimental PIN data bases one records only whether or not protein pairs interact, not the *number* of complexes in which they interact. Hence, protein interactions are normally represented in terms of the adjacency matrix  $\mathbf{a} = \{a_{ij}\}$ , which is related to the weighted matrix  $\mathbf{c} = \{c_{ij}\}$  via  $a_{ij} = \theta(c_{ij}) \forall (i \neq j)$ , with the convention for the step function  $\theta(0) = 0$ . We therefore have  $p(a_{ij}) = \langle \delta_{c_{ij},0} \rangle \delta_{a_{ij},0} + (1 - \langle \delta_{c_{ij},0} \rangle) \delta_{a_{ij},1}$ . However, the links  $\{a_{ij}\}$  are correlated. In Appendix C we derive the relation between the expected values of different graph observables for the two graph ensembles  $p(\mathbf{a})$  and  $p(\mathbf{c})$ . Denoting averages in the  $\mathbf{a}$  ensemble as  $\langle \dots \rangle_a$ , and using the usual notation  $\langle \dots \rangle$  for averages in the  $\mathbf{c}$  ensemble, one finds that for large  $N$  the first two moments of the degree distributions and the first two loop densities in the two ensembles are identical:

$$\begin{aligned} \langle k \rangle_a &= \frac{1}{N} \sum_{ij} \langle a_{ij} \rangle_a = \frac{1}{N} \sum_{ij} [1 - \langle \delta_{c_{ij},0} \rangle] = \alpha \langle q^2 \rangle + \mathcal{O}(N^{-1}) \\ &= \langle k \rangle + \mathcal{O}(N^{-1}) \end{aligned} \quad (42)$$

$$\begin{aligned} \langle k^2 \rangle_a &= \frac{1}{N} \sum_{i \neq j \neq k} \langle a_{ij} a_{jk} \rangle = \alpha \langle q^2 \rangle + \alpha \langle q^3 \rangle + \alpha^2 \langle q^2 \rangle^2 + \mathcal{O}(N^{-1}) \\ &= \langle k^2 \rangle + \mathcal{O}(N^{-1}) \end{aligned} \quad (43)$$

$$\begin{aligned} m_3^a &= \frac{1}{N} \sum_{i \neq j \neq k (\neq i)} \langle a_{ij} a_{jk} a_{ki} \rangle = \alpha \langle q^3 \rangle + \mathcal{O}(N^{-1}) \\ &= m_3 + \mathcal{O}(N^{-1}) \end{aligned} \quad (44)$$

$$\begin{aligned} m_4^a &= \frac{1}{N} \sum_{[i,j,k,\ell]} \langle a_{ij} a_{jk} a_{k\ell} a_{\ell i} \rangle = \alpha \langle q^4 \rangle + \mathcal{O}(N^{-1}) \\ &= m_4 + \mathcal{O}(N^{-1}) \end{aligned} \quad (45)$$

Square brackets underneath summations again indicates distinct indices, which excludes backtracking in the counting of length-4 loops. Apparently, the ensembles  $p(\mathbf{a})$  and  $p(\mathbf{c})$  are asymptotically equivalent with regard to the statistics of these four quantities. We will see in the next section that this equivalence holds also for the ‘dual’ ensemble (3). To test the above claims we compute and show in Figure 3 the above observables in synthetic graphs  $\mathbf{c}$  and  $\mathbf{a}$  generated randomly from (10,11), where the random bipartite interaction graph  $\mathbf{\xi}$  is drawn from (1).

#### 4. Network properties generated by the $d$ -ensemble

In this section we will derive properties for the network ensembles (10,11) upon assuming that the statistics of the underlying bipartite protein interaction network are given by (3), i.e. are protein-driven as opposed to complex-driven. In spite of the superficial similarity between definitions (2) and (4), the expectations of graph observables in the two ensembles are found to be remarkably different.

##### 4.1. Link probabilities

We start by calculating the link expectation values in the weighted graphs  $c_{ij} = \sum_{\mu} \xi_i^{\mu} \xi_j^{\mu}$ :

$$\langle c_{ij} \rangle = \sum_{\mu} \langle \xi_i^{\mu} \xi_j^{\mu} \rangle = \frac{d_i d_j}{\alpha N} \quad (46)$$

Hence the random graphs  $\mathbf{c}$  are again finitely connected, now with

$$\langle k \rangle = \frac{1}{N} \sum_{ij} \langle c_{ij} \rangle = \frac{\langle d \rangle^2}{\alpha} \quad (47)$$

Averages over  $d$  refer to the distribution  $P(d)$  of protein promiscuities in the bipartite graph  $\mathbf{\xi}$ . The result (47) can also be written as  $\langle k \rangle = \alpha \langle q \rangle^2$ , and is thus notably different from the earlier expression  $\langle k \rangle = \alpha \langle q^2 \rangle$  found in the  $q$ -ensemble. The link likelihood is calculated in Appendix A, and shows again that  $p(c_{ij} > 1) = \mathcal{O}(N^{-2})$ .

##### 4.2. Densities of short loops

We can calculate the density of length-3 loops similar to how this was done for the  $q$ -ensemble in the previous section. Again these are given, to order  $\mathcal{O}(1)$ , by the  $S_3$  stars in the bi-partite graph, since the contribution from combinations of  $S_2$  stars is as before  $\mathcal{O}(N^{-1})$ . Here we obtain

$$m_3 = \frac{1}{N} \sum_{[ijk]} \sum_{\mu} \langle \xi_i^{\mu} \xi_j^{\mu} \xi_k^{\mu} \rangle = \frac{1}{N} \sum_{[ijk]} \sum_{\mu} \frac{d_i d_j d_k}{\alpha^3 N^3} = \frac{\langle d \rangle^3}{\alpha^2} \quad (48)$$

For loops of arbitrary length  $L$  this generalises to

$$m_L = \langle d \rangle^L / \alpha^{L-1} \quad (49)$$

Interestingly, the densities  $m_L$  of short loops and the average connectivity  $\langle k \rangle$  depend on  $P(d)$  only through its first moment. Promiscuity heterogeneity apparently cannot affect the densities of short loops. In the present ensemble these densities must therefore be identical to what would be found in a randomly wired bipartite graph. This prediction will be confirmed in simulations.

### 4.3. The degree distribution

In Appendix B we calculate the asymptotic degree distribution of  $\mathbf{c}$  for the protein-driven complex recruitment model (3), giving

$$p(k) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_i \delta_{k, \sum_j c_{ij}} = \sum_{d \geq 0} P(d) \sum_{\ell} \left( e^{-d} d^{\ell} / \ell! \right) \left( e^{-\ell \frac{\langle d \rangle}{\alpha}} \left( \frac{\ell \langle d \rangle}{\alpha} \right)^k / k! \right) \quad (50)$$

This result is again understood easily: the number of neighbours of a node  $i$  is a Poissonian variable  $\ell$ , with average  $d$ , where  $d$  is now drawn from  $P(d)$ . Each of the  $\ell$  first neighbours will have a degree which is a Poissonian variable with average  $\langle d \rangle / \alpha$ , so the number  $k$  of second neighbours of  $i$  in the bipartite graph is a Poisson variable with average  $\ell \langle d \rangle / \alpha$ . Equation (50) shows that a tail in the promiscuity distribution  $P(d)$  will induce a tail in the degree distribution  $p(k)$  of  $\mathbf{c}$ . The link between the two distributions is again most easily expressed via generating functions. Upon defining  $Q_1(z) = \sum_k p(k) e^{-zk}$  and  $Q_4(z) = \sum_d P(d) e^{-zd}$ , we obtain from (50):

$$Q_1(z) = \sum_{d \geq 0} P(d) e^{-d} \sum_{\ell} \left( d e^{\langle d \rangle (e^{-z} - 1) / \alpha} \right)^{\ell} / \ell! = Q_4(1 - e^{\langle d \rangle (e^{-z} - 1) / \alpha}) \quad (51)$$

For  $z \simeq 0$  this gives

$$Q_1(z) \simeq Q_4(z \langle d \rangle / \alpha) \quad (52)$$

Hence, if  $p(k)$  decays for large  $k$  as  $p(k) \simeq C k^{-\mu}$  with  $2 < \mu < 3$ , then via (32) we infer that

$$Q_4(z \langle d \rangle / \alpha) \simeq 1 - \langle k \rangle z + C \Gamma(1 - \mu) z^{\mu-1} \quad (53)$$

Equivalently,

$$Q_4(x) \simeq 1 - \alpha \langle k \rangle x / \langle d \rangle + C \Gamma(1 - \mu) (\alpha / \langle d \rangle)^{\mu-1} x^{\mu-1} \quad (54)$$

This implies that for large  $d$  the promiscuity distribution will be of the form  $P(d) \simeq C' d^{-\mu}$ , where

$$C' = C (\alpha / \langle d \rangle)^{\mu-1} = C \langle q \rangle^{1-\mu} \quad (55)$$

Any tail in the promiscuity distribution will produce the same tail in the degree distribution of  $\mathbf{c}$ , but with a rescaled amplitude. Fat tails in the degree distribution of protein interaction networks can thus arise from equally heterogeneous ‘dating’ interactions between proteins, combined with a homogeneous distribution of ‘party’ interactions. Short loops are boosted by broad distributions of complex sizes, since large complexes in the bipartite graph induce large cliques in the network  $\mathbf{c}$ . The  $d$ -ensemble (3), which attributes any heterogeneity in  $p(k)$  to heterogeneity of protein binding promiscuities, generates separable PIN graphs  $\mathbf{c}$  with the least number of loops. Conversely, the  $q$ -ensemble (1), which attributes all heterogeneity in  $p(k)$  to heterogeneity in complex sizes, generates separable PIN graphs  $\mathbf{c}$  with the largest number of loops.

### 4.4. Relations that are independent of $P(d)$ and $\alpha$

The first two moments of the degree distribution  $p(k)$  of the separable PIN networks  $\mathbf{c}$  are

$$\langle k \rangle = \sum_k k p(k) = \sum_d P(d) \sum_{\ell} e^{-d} \frac{d^{\ell}}{\ell!} \frac{\ell \langle d \rangle}{\alpha} = \langle d \rangle^2 / \alpha \quad (56)$$

$$\begin{aligned}\langle k^2 \rangle &= \sum_k k^2 p(k) = \sum_d P(d) \sum_\ell e^{-d} \frac{d^\ell}{\ell!} \left[ \left( \frac{\ell \langle d \rangle}{\alpha} \right)^2 + \frac{\ell \langle d \rangle}{\alpha} \right] \\ &= \langle d \rangle^2 / \alpha + \langle d \rangle^3 / \alpha^2 + \langle d \rangle^2 \langle d^2 \rangle / \alpha^2\end{aligned}\quad (57)$$

Combination of (63), (57) and (48) now yields the relation

$$\langle d^2 \rangle / \alpha = (\langle k^2 \rangle - \langle k \rangle - m_3) / \langle k \rangle \quad (58)$$

which still involves  $\langle d^2 \rangle$  and  $\alpha$ . We can also find an alternative expression for the density of loops of length 3 by combining (63) and (48)

$$m_3 = \langle k \rangle^{3/2} / \sqrt{\alpha} \quad (59)$$

Unfortunately, neither of our two expressions for  $m_3$ , (58) nor (59), are useful, because the protein promiscuities distribution  $P(d)$  and the ratio  $\alpha$  are generally unknown. Access to information on these quantities via future detection experiments may therefore be extremely welcome in support of theoretical modelling of protein interaction datasets. To make progress, we need to derive relations for graph observables that are independent of  $\alpha$  and  $P(d)$ . We note that (49) yields

$$\forall L \geq 3: \quad m_{L+1} / m_L = \langle d \rangle / \alpha \quad (60)$$

This can be rewritten using (63), as

$$\forall L \geq 3: \quad m_{L+1} / m_L = \sqrt{\langle k \rangle / \alpha} \quad (61)$$

On the other hand, we know from (59) that  $m_3 / \langle k \rangle = \sqrt{\langle k \rangle / \alpha}$ . Combining the above formulae allows us to establish the following relation, that now is completely independent of  $P(d)$  and  $\alpha$ :

$$m_4 = m_3^2 / \langle k \rangle \quad (62)$$

Again we have tested the various formulae in synthetically generated graphs, see Figure 4.

#### 4.5. Link between **a** and **c** graph definitions

As a final step, we check whether the observables  $m_3$  and  $m_4$  are indeed the same for the two PIN definitions (10, 11), with the bipartite graph of our protein-driven ensemble (3), since protein detection experiments provide the binary matrix **a** as opposed to the weighted graph **c** for which (66) was derived. Again we denote averages relating to **a** as  $\langle \dots \rangle_a$ , and those relating to **c** as  $\langle \dots \rangle$ . For the moments of the degree distributions we find the differences to be negligible:

$$\langle k \rangle_a = \frac{1}{N} \sum_{ij} \langle a_{ij} \rangle_a = \frac{\langle d \rangle^2}{\alpha} + \mathcal{O}(N^{-1}) = \langle k \rangle + \mathcal{O}(N^{-1}) \quad (63)$$

$$\langle k^2 \rangle_a = \frac{1}{N} \sum_{i \neq j \neq k} \langle a_{ij} a_{jk} \rangle = \frac{\langle d \rangle^2}{\alpha} + \frac{\langle d \rangle^3}{\alpha^2} + \frac{\langle d^2 \rangle \langle d \rangle^2}{\alpha^2} + \mathcal{O}(N^{-1}) = \langle k^2 \rangle + \mathcal{O}(N^{-1}) \quad (64)$$

The same is true for the densities of loops of length 3 and 4:

$$m_3^a = \frac{1}{N} \sum_{i \neq j \neq k (\neq i)} \langle a_{ij} a_{jk} a_{ki} \rangle = \frac{\langle d \rangle^3}{\alpha^2} + \mathcal{O}(N^{-1}) = m_3 + \mathcal{O}(N^{-1}) \quad (65)$$

$$m_4^a = \frac{1}{N} \sum_{[i,j,k,\ell]} \langle a_{ij} a_{jk} a_{k\ell} a_{\ell i} \rangle = \frac{\langle d \rangle^4}{\alpha^3} + \mathcal{O}(N^{-1}) = m_4 + \mathcal{O}(N^{-1}) \quad (66)$$

This equivalence between the ensembles  $p(\mathbf{a})$  and  $p(\mathbf{c})$  when calculating the main average values of graph observables for large  $N$  implies that large protein interaction adjacency matrices can in practice be regarded as having a separable structure. Again, we check our relations (63, 57, 65, 66), against synthetically generated graphs and show results in figure 4.5.

## 5. Macroscopic observables in the mixed ensemble

The two bipartite graph ensembles (1, 3) considered so far led to Poissonian distributions either for the protein promiscuities  $d_i$  (in the  $q$ -ensemble), or for the complex sizes  $q_\mu$  (in the  $d$  ensemble). It is possible to model heterogeneity in both  $d_i$  and  $q_\mu$  using the mixed ensemble (5). Due to the similarities with previous calculations we can and will be more brief in this section. For ensemble (5) the expectation values of individual links in the weighted graph  $\mathbf{c}$  are

$$\langle c_{ij} \rangle = \sum_{\mu} \langle \xi_i^{\mu} \xi_j^{\mu} \rangle = \sum_{\mu} \frac{d_i d_j q_{\mu}^2}{\alpha^2 \langle q \rangle^2 N^2} = \frac{d_i d_j \langle q^2 \rangle}{\alpha \langle q \rangle^2 N} + \mathcal{O}(N^{-3/2}) \quad (67)$$

and the average connectivity follows as

$$\langle k \rangle = \frac{1}{N} \sum_{ij} \langle c_{ij} \rangle = \frac{\langle d \rangle^2 \langle q^2 \rangle}{\alpha \langle q \rangle^2} + \mathcal{O}(N^{-1/2}) = \alpha \langle q^2 \rangle + \mathcal{O}(N^{-1/2}) \quad (68)$$

Full details are found in Appendix A. As in previous ensembles, the leading contribution to the density of length-3 loops comes from the  $S_3$  stars in the bipartite graphs, now giving

$$m_3 = \frac{1}{N} \sum_{[ijk]} \sum_{\mu} \langle \xi_i^{\mu} \xi_j^{\mu} \xi_k^{\mu} \rangle = \frac{1}{N} \sum_{[ijk]} \sum_{\mu} \frac{d_i d_j d_k q_{\mu}^3}{\alpha^3 \langle q \rangle^3 N^3} \simeq \frac{\langle d \rangle^3 \langle q^3 \rangle}{\alpha^2 \langle q \rangle^3} = \alpha \langle q^3 \rangle \quad (69)$$

As before, the heterogeneity in the  $d$  affects neither the average connectivity  $\langle k \rangle$  nor the density of triangles  $m_3$ , both are as they were in the  $q$ -ensemble. This is confirmed numerically, see Figure 6. The degree distribution for large  $N$  in the ensemble  $p(\mathbf{c})$  is calculated in Appendix B, giving

$$p(k) = \int_0^{\infty} dy P(y) e^{-y} y^k / k! \quad (70)$$

where

$$P(y) = \sum_d P(d) e^{-d} \sum_{\ell \geq 0} \frac{d^{\ell}}{\ell!} \sum_{q_1 \dots q_{\ell} \geq 0} W(q_1) \dots W(q_{\ell}) \delta[y - \sum_{r \leq \ell} q_r] \quad (71)$$

Again it is possible to relate the asymptotic behaviour of  $p(k)$  to that of  $P(d)$  and  $W(q)$ , by inspecting the relation between the relevant generating functions. Using our previous definitions for  $Q_1(z)$ ,  $Q_2(z)$ ,  $Q_3(z)$ , and  $Q_4(z)$ , we obtain via (70) and (71):

$$\begin{aligned} Q_1(z) &= \int dy P(y) \sum_k e^{-y} (y e^{-z})^k / k! = \int dy P(y) e^{-y(1-e^{-z})} \\ &= Q_2(1 - e^{-z}) \end{aligned} \quad (72)$$

$$\begin{aligned} Q_2(z) &= \sum_d P(d) e^{-d} \sum_{\ell} \frac{d^{\ell}}{\ell!} \prod_{r=1}^{\ell} \left( \sum_{q_r} W(q_r) e^{-z q_r} \right) = \sum_d P(d) e^{-d} \sum_{\ell} \frac{d^{\ell}}{\ell!} Q_3^{\ell}(z) \\ &= \sum_d P(d) e^{-d[1-Q_3(z)]} = Q_4(1 - Q_3(z)) \end{aligned} \quad (73)$$

Expanding (72) for small  $z$  tells us that  $Q_1(z) \simeq Q_2(z)$ . Substitution into (73) subsequently gives

$$Q_1(z) \simeq Q_4(1 - Q_3(z)) \quad (74)$$

Assuming  $W(q)$  to have a power-law tail, but with a finite first moment (as in all cases previously considered), i.e.  $W(q) \simeq Kq^{-\gamma}$  with  $\gamma > 2$ , its generating function  $Q_3(z)$  can be written as

$$Q_3(z) = 1 - \langle q^2 \rangle z / \langle q \rangle + \mathcal{O}(z^\delta) \quad (75)$$

where  $\delta = \min\{2, \gamma - 1\}$ . Insertion into (74) then leads to

$$Q_1(z) \simeq Q_4(z \langle q^2 \rangle / \langle q \rangle - \mathcal{O}(z^\delta)) \quad (76)$$

If  $p(k) = Ck^{-\mu}$ , with  $2 < \mu < 3$ , we may use our earlier result (32) and get

$$Q_4(x - \mathcal{O}(x \langle q \rangle / \langle q^2 \rangle)^\delta) \simeq 1 - \langle k \rangle \langle q \rangle x / \langle q^2 \rangle + C\Gamma(1-\mu)(\langle q \rangle / \langle q^2 \rangle)^{\mu-1} x^{\mu-1} \quad (77)$$

If  $\gamma > \mu$  we have  $\delta > \mu - 1$ , so we can neglect the second term in the argument of  $Q_4$  and conclude that the promiscuity distribution has the asymptotic form  $P(d) = C'd^{-\mu}$  where  $C' = C(\langle q^2 \rangle / \langle q \rangle)^{1-\mu}$ . This means that if  $W(q)$  decays faster than  $p(k)$  (as in Section 4), then the tail in  $p(k)$  must arise from the tail in  $P(d)$ . Note, however, that heterogeneities in  $P(q)$  will affect the amplitude of the power law tail in  $P(d)$ , which will be smaller by a factor  $(\langle q^2 \rangle / \langle q \rangle)^{1-\mu}$  compared to the case where  $P(q) = \delta_{q,\langle q \rangle}$ , where we had  $C' = C\langle q \rangle^{1-\mu}$ . Conversely, if  $\gamma = \mu$  we have  $\delta = \mu - 1$ , and writing the  $\mathcal{O}(z^\delta)$  term explicitly in (76) gives

$$Q_4(z \langle q^2 \rangle / \langle q \rangle - K\Gamma(1-\mu)z^{\mu-1}) = 1 - \langle k \rangle z + C\Gamma(1-\mu)z^{\mu-1} \quad (78)$$

Expanding both sides in powers of  $z$  and equating prefactors tells us that either  $C' = 0$  and  $C = K\langle d \rangle$  (i.e.  $K = C/\alpha\langle q \rangle$ , which retrieves the case in Section 3), or  $\delta = \mu$  with  $K\langle d \rangle + C'(\langle q^2 \rangle / \langle q \rangle)^{\mu-1} = C$ . Hence, if  $P(d)$  is as broad as  $W(q)$ , then both contribute to the tail in  $p(k)$ , whose amplitude will be the sum of the amplitudes of the tails in  $P(q)$  and  $P(d)$ . We see in (77) that  $\gamma < \mu$  is not possible, i.e.  $W(q)$  needs to decay at least as fast as  $p(k)$ .

In Appendix B we calculate the first two moments of the degree distribution  $p(k)$  of the ensemble  $p(\mathbf{c})$ . This recovers (68) for the first moment, and for the second moment gives

$$\langle k^2 \rangle = \alpha \langle q^2 \rangle + \alpha \langle q^3 \rangle + \langle d^2 \rangle \langle k \rangle^2 / \langle d \rangle^2 \quad (79)$$

Substituting (68) and (69) into (79) then leads to

$$m_3 = \langle k^2 \rangle - \langle k \rangle - \langle k \rangle^2 \langle d^2 \rangle / \langle d \rangle^2 \quad (80)$$

The density of length-3 loops depends again on the first two moments of the degree distribution  $p(k)$ , but is also seen to depend on the first two moments of the promiscuity distribution  $P(d)$ , which is unknown. Hence, this relation cannot serve as a test of PIN data quality. It is nevertheless useful for comparing the mixed ensemble to the  $d$ - and the  $q$ -ensembles in synthetically generated data.

## 6. Numerical comparison of the three bipartite generative ensembles

Here we compare the ability of our bipartite ensembles (1, 3, 5) to predict properties of the associated binary PIN graphs, for synthetic networks that are generated from any of these ensembles. We focus on comparing homologous formulae for the observables  $\langle k \rangle$ ,  $\langle k^2 \rangle$ ,  $m_3$  and  $m_4$ . The synthetic matrices  $\mathbf{a} = \{a_{ij}\}$  with  $a_{ij} \in \{0, 1\}$  are defined as before via  $a_{ij} = \theta(\sum_{\mu} \xi_i^{\mu} \xi_j^{\mu})$ , with  $\theta(0) = 0$ , and the links of the bipartite graph  $\xi$  are generated from the following three protocols. In the first protocol, links between nodes  $(i, \mu)$  are drawn randomly and independently, until their total number reaches a prescribed limit. In the second protocol, we assign the links preferentially to complexes with large sizes. In a third protocol we assign links preferentially to proteins with large promiscuities.

In Figure 6 we show along the vertical axes the values of  $\langle k \rangle$  (left) predicted by the three ensembles, via formulae (37), (47) and (68), the predicted values of  $\langle k^2 \rangle$  (middle), via (38), (57), and (79), and the predicted triangle density  $m_3$  (right), via (40), (58) and (80). All are shown together with the corresponding values that were measured in  $\mathbf{a}$ , along the horizontal axis. As expected, the  $d$ -ensemble outperforms the other ensembles when links are drawn according to  $d$ -preferential attachment, whereas the  $q$ -ensemble performs better for graphs generated via  $q$ -preferential attachment. The mixed ensemble performs very similar to the  $q$ -ensemble in terms of counting triangles, as expected from the reasoning in Section 5. Deviations between the  $q$  and the mixed ensembles are most evident in the second moment of the degree distribution, where the mixed ensemble always leads to values well above those of the  $q$ - and the  $d$ -ensembles. We found in Section 4 that the  $d$ -ensemble is indistinguishable from a fully random ensemble when calculating  $\langle k \rangle$  and  $m_3$ , which explains why the  $d$ -ensemble predicts the values of these two observables perfectly. The other two ensembles are more sensitive to finite size effects, as any heterogeneity in the  $q$  will boost the number of loops.

In Figure 7 we show the values of  $m_3$  and  $m_4$  predicted by those formulae that involve only measurable graph observables, for the synthetically generated graphs used in Figure 6. The prediction of  $m_3$  is now obtained from (40) and (66), for the  $q$ - and  $d$ -ensembles respectively, and  $m_4$  is evaluated using (41) and (66). In figure 8 we plot the degree distribution  $p(k)$  of graphs with identical values for the number of nodes ( $N = 3000$ ) and the number of links  $L = N\alpha\langle q \rangle$ , generated synthetically via the three chosen protocols, together with the distributions  $P(q)$  of complex sizes and  $P(d)$  of protein promiscuities. As explained in Section 5, tails in the degree distribution  $p(k) \sim k^{-\mu}$  can arise either from a complex size distribution  $P(q) \sim q^{-\mu-1}$  and a homogeneous promiscuity distribution, or from having an equally fat tail in the promiscuity distribution  $P(d) \sim d^{-\mu}$  together with less heterogeneous complex sizes  $P(q) \sim q^{-\alpha-1}$  with  $\alpha > \mu$ .

## 7. Test against experimental protein interaction data

In this section we apply the results of our analyses to real publicly available protein interaction datasets, obtained via MS (mass spectrometry) and Y2H (yeast 2-hybrid) experiments. The detailed quantitative features of the various data sets and their references are listed in Table 7.



Species	$N$	$\langle k \rangle$	$k_{\max}$	Method	Reference
<i>C.elegans</i>	2528	2.96	99	Y2H	[25]
<i>C.jejuni</i>	1324	17.5	207	Y2H	[26]
<i>E.coli</i>	2457	7.05	641	MS	[24]
<i>H.pylori</i>	724	3.87	55	Y2H	[27]
<i>H.sapiens</i> I	1499	3.37	125	Y2H	[28]
<i>H.sapiens</i> II	1655	3.71	95	Y2H	[29]
<i>H.sapiens</i> III	2268	5.67	314	MS	[30]
<i>M.loti</i>	1803	3.43	401	Y2H	[31]
<i>P.falciparum</i>	1267	4.17	51	Y2H	[32]
<i>S.cerevisiae</i> I	991	1.82	24	Y2hH	[33]
<i>S.cerevisiae</i> II	787	1.91	55	Y2H	[34]
<i>S.cerevisiae</i> III	3241	2.69	279	Y2H	[34]
<i>S.cerevisiae</i> IV	1576	4.58	62	MS	[35]
<i>S.cerevisiae</i> VI	1358	4.73	53	MS	[36]
<i>S.cerevisiae</i> VIII	2551	16.77	955	MS	[37]
<i>S.cerevisiae</i> IX	2708	5.25	141	MS	[38]
<i>Synechocystis</i>	1903	3.25	51	Y2H	[39]
<i>T.pallidum</i>	724	10.01	285	Y2H	[40]

**Table 1.** List of the publicly available experimental protein interaction data sets as used in the present study, together with their main quantitative characteristics (number of proteins  $N$ , average degree  $\langle k \rangle$ , and largest degree  $k_{\max}$ ) and references.

### 7.1. Mass spectrometry datasets

Seven of the experimental PIN datasets in Table 7 were obtained by MS experiments, and they involved three distinct biological species, namely *S. cerevisiae*, *H.sapiens* and *E.coli*. Each set takes the form of an  $N \times N$  matrix of binary entries  $a_{ij}$ , but with different values of  $N$ .

In Figure 9 we show the results of our analytical predictions for the densities of length-3 and length-4 loops, as given by the formulae for the bipartite  $q$ - and  $d$ -ensembles, versus their measured values in the MS datasets. The  $q$ -ensemble leads to values of the number of short loops consistently higher than those predicted by the  $d$ -ensemble. This could have been expected, since the  $q$ -ensemble induces large cliques in the protein interaction networks  $\mathbf{c}$  and  $\mathbf{a}$ , which boosts short loops. In contrast, the  $d$ -ensemble induces a homogeneous distribution for the complex sizes, and thereby suppresses the presence of large cliques in the protein interaction networks.

Remarkably, the values for length-4 loop densities of all the MS data sets are in between those of the  $d$ -ensemble (which thereby acts as a lower bound) and those of the  $q$ -ensemble (which acts as an upper bound). This suggests a compatibility of data from MS experiments with the expected separable form of the proteome network. However, the measured length-3 densities are consistently lower than the values compatible with a separable structure of the proteome.

## 7.2. Yeast 2-hybrid datasets

We tested similarly the compatibility of Y2H data with a separable structure of the proteome, by checking whether the measured values for the network observables  $m_3$  and  $m_4$  fall within what appeared to be (in MS data) theoretical bounds set by the  $q$ - and  $d$ -ensembles. We now used the 12 PIN datasets in Table 7 that were obtained from Y2H experiments. Results are shown in Figure 10. We observe that Y2H datasets exhibit generally fewer short loops than MS dataset. This may be due to the fact that Y2H experiments mostly detect direct binding domain contacts in protein interactions, leading to an undersampling of links (and thereby to an underestimation of connectivity and loops). However, Y2H data sets still show the same level of compatibility with a separable structure of the proteome as the MS datasets did, with measured values of  $m_4$  that are fully compatible, and values for  $m_3$  that fall below those predicted by the  $d$ -ensemble. This is quite remarkable, since MS and Y2H experiments are known to measure interactions in very different ways.

## 8. Conclusions

In this paper we propose a bipartite network representation of protein interactions, where the two node types represent proteins and complexes, respectively. A protein-protein interaction network can then be regarded as the result of a ‘marginalization’ of the bipartite network, whereby the complexes are integrated out (i.e. summed over). This leads to a weighted protein interaction network  $\mathbf{c}$  with a separable structure. Adjacency matrices of protein interaction networks  $\mathbf{a}$  are then simply the binary versions of the separable  $\mathbf{c}$ , obtained by the entry truncations  $a_{ij} = \theta(c_{ij})$ , with the convention  $\theta(0) = 0$ . One of the central results of this work is that for sufficiently large networks there is an equivalence between the two graph ensembles  $p(\mathbf{c})$  and  $p(\mathbf{a})$ , inasmuch as macroscopic statistical properties are concerned, such as densities of short loops and degree distributions. This allows us to regard the conventional protein interaction adjacency matrices as if they were to have a separable structure, and induces precise relations between expectation values of macroscopic graph observables which, remarkably, only depend on measurable quantities and on the underlying mechanism with which proteins and complexes recruit each other. They are independent of inaccessible microscopic details of proteins and their complexes.

We considered the two extreme complex recruitment scenarios, one where recruitment is either driven solely by protein promiscuities, and one where it is driven by complex sizes. Preferential attachment to large complexes (the  $q$ -ensemble) favours the presence of large cliques in PINs, which boosts the number of short loops. Hence we can reasonably expect that the predictions on short loop densities from the  $q$ -ensemble will over-estimate the real number of loops. Conversely, preferential attachment based only on protein promiscuities (the  $d$ -ensemble) leads to homogeneous complex sizes, which suppresses large cliques in PINs, leading to an underestimation of short loop densities. Remarkably, real protein interaction data from mass-spectrometry and yeast 2-hybrid experiments show a density of length-4 loops in between the predictions of the  $d$ -ensemble and those of the  $q$ -ensemble, suggesting a degree of compatibility of these experimental data with a separable structure of the proteome. In contrast, both MS and Y2H dataset show

densities or length-3 loops that are consistently smaller than all our theoretical predictions.

We believe that, by providing a systematic and practical framework for understanding protein interaction experiments, our approach may represent a valuable step towards establishing a more solid connection between protein interaction datasets and the underlying biology. Universal bounds on observables in PINs may become powerful tools for data quality testing. Improved versions of the present models, with fit the experimental data better, may open a route to infer quantities such as the ratio  $\alpha$ , and the distributions of protein promiscuities and complex sizes. Such quantities are not available in the current PIN data sets, and are difficult to access experimentally. The present work has revealed that the asymptotic forms of these distributions can be extracted from the tails of the PIN degree distributions. Finally, our method may shed some light on the way protein and complexes recruit one another, in particular, whether this recruitment is driven by proteins or by complexes, and may enable us to discriminate between ‘party hub’ and ‘date hub’ interactions.

## 9. Acknowledgements

AA acknowledges Alessandro Pandini and Sun Chung for providing protein interaction datasets. Kate Roberts is acknowledged for interesting discussions during the early stages of this work. ACCC is grateful for support from the UK’s Biotechnology and Biological Sciences Research Council (BBSRC) .

## 10. References

- [1] Hakes L, Pinney JW, Robertson DL, Lovel SC 2008 *Nature Biotechnology* **26** 69-72
- [2] Han JDJ, Dupuy D, Bertin N, Cusick ME and Vidal M 2005 *Nature Biotechnology* **23**, 839–844.
- [3] De Silva E, Thorne T, Ingram P, Agrafioti I, Swire J, Wiuf C and Stumpf MPH 2006 *BMC Biology* **4**, 39.
- [4] Fernandes LP, Annibale A, Kleinjung J, Coolen ACC and Fraternali F 2010 *PLoS ONE* **5**, e12083.
- [5] Lee SH, Kim PJ and Jeong H 2006 *Phys. Rev. E* **73**, 016102.
- [6] Stumpf MPH and Wiuf C 2005 *Phys. Rev. E* **72**, 036118.
- [7] Stumpf MPH, Wiuf C and May RM 2005 *Proc. Natl. Sci. USA* **102**, 4221–4224.
- [8] Viger F, Barrat A, Dall’Asta L, Zhang CH and Kolaczyk ED 2007 *Phys. Rev. E* **75**, 056111.
- [9] Solokov IM and Eliazar II 2010 *Phys. Rev. E* **81**, 026107.
- [10] Annibale A, Coolen ACC 2011 *Interface Focus* **2011** **1**(6) .
- [11] Newman, Strogatz, Watts *Phys. Rev. E* **64** 026118
- [12] E Agliari, A Annibale, A Barra, ACC Coolen, D Tantari J. Phys. A: Math. Theor: **46** (41) 2013
- [13] P Sollich, D Tantari, A Annibale, A Barra preprint *arXiv:1404.3654*
- [14] P Sollich, D Tantari, A Annibale, A Barra preprint *arXiv:1404.3654*
- [15] ES Roberts and ACC Coolen
- [16] Maslov S and Sneppen K 2002 *Science* **296** 5569
- [17] Chang X, Xu T, Li Y and Wang K *Scientific Reports* **3** 1691
- [18] Albert R and Barabassi A L 2002 *Reviews of Modern Physics* **74** 47-97
- [19] Barabasi A L and Albert R 1999 *Science* **286** 509
- [20] Dorogovtsev S N and Mendes J F 2003 *Evolution of networks* (Oxford: Oxford University Press)
- [21] Abramowitz M and Stegun IA 1972 *Handbook of mathematical functions* (New York: Dover)
- [22] Junker B H and Schreiber F 2008 *Analysis of biological networks* (New York: Wiley Series on Bioinformatics)
- [23] Ivanic J, Wallqvist A and Reifman J 2008 *PLoS Computational Biology* **4** 7

- [24] Arifuzzaman M *et al.* 2006 *Genome Res* **16** 686-691
- [25] Simonis N *et al.* 2008 *Nature Methods* **6**(1):47-54
- [26] Parrish J R *et al.* 2007 *Genome Biology* **8**(7)
- [27] Rain J C *et al.* 2001 *Nature* **409**(6817):211-215
- [28] Rual J-F F *et al.* 2005 *Nature* **437**(7062):1173-1178
- [29] Stelzl U *et al.* 2005 *Cell* **122**(6):957-968
- [30] Ewing R M *et al.* 2007 *Molecular systems biology* **3**:89
- [31] Shimoda Y, Shinpo S, Kohara M, Nakamura Y, Tabata S and Sato S 2008 *DNA Res* **15**(1):13-23
- [32] Lacount D J *et al.* 2005 *Nature* **438**(7064):103-107
- [33] Uetz P *et al.* 2000 *Nature* **403**(6770):623-627
- [34] Ito T *et al.* 2001 *Proc Natl Acad Sci U S A* **98**(8):4569-4574
- [35] Ho Y *et al.* 2002 *Nature* **415**(6868):180-183
- [36] Gavin A C *et al.* 2002 *Nature* **415**(6868):141-147
- [37] Gavin A C C *et al.* 2006 *Nature* **440**(7084):631-636
- [38] Krogan N J *et al.* 2006 *Nature* **440**(7084):637-643
- [39] Sato S, Shimoda Y, Muraki A, Kohara M, Nakamura Y and Tabata S 2007 *DNA Res* **14**(5), 207-216
- [40] Titz B *et al.* 2008 *PLoS ONE* **3**(5)

## Appendix A. Link probabilities in the weighted protein interaction network

In this appendix we derive the likelihood to have a link in the weighted protein interaction network  $c_{ij} = \sum_{\mu} \xi_i^{\mu} \xi_j^{\mu}$ , when the  $\xi_i^{\mu}$  are drawn from the ensembles (1,3,5).

### Appendix A.1. The $q$ ensemble

In the  $q$ -ensemble we have

$$\begin{aligned}
 p(c_{ij}) &= \left\langle \delta_{c_{ij}, \sum_{\mu \leq \alpha N} \xi_i^{\mu} \xi_j^{\mu}} \right\rangle_{\xi} = \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega c_{ij}} \prod_{\mu=1}^{\alpha N} \left\langle e^{-i\omega \xi_i^{\mu} \xi_j^{\mu}} \right\rangle_{\xi} \\
 &= \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega c_{ij}} \prod_{\mu=1}^{\alpha N} \left\{ \frac{q_{\mu}^2}{N^2} e^{-i\omega} + \left(1 - \frac{q_{\mu}^2}{N^2}\right) \right\} = \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega c_{ij} + \sum_{\mu=1}^{\alpha N} \frac{q_{\mu}^2}{N^2} [e^{-i\omega} - 1] - \frac{1}{2} \sum_{\mu=1}^{\alpha N} \frac{q_{\mu}^4}{N^4} [e^{-i\omega} - 1]^2} \\
 &= \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega c_{ij}} \left[ 1 + \frac{\alpha \langle q^2 \rangle}{N} (e^{-i\omega} - 1) - \frac{1}{2} \frac{\alpha \langle q^4 \rangle}{N^3} (e^{-i\omega} - 1)^2 + \frac{\alpha^2 \langle q^2 \rangle^2}{2N^2} (e^{-i\omega} - 1)^2 \right. \\
 &\quad \left. + \frac{\alpha^3 \langle q^2 \rangle^3}{6N^3} (e^{-i\omega} - 1)^3 + \mathcal{O}(N^{-4}) \right] \\
 &= \delta_{c_{ij},0} + \frac{\alpha \langle q^2 \rangle}{N} (\delta_{c_{ij},1} - \delta_{c_{ij},0}) + \left( \frac{\alpha^2 \langle q^2 \rangle^2}{2N^2} - \frac{1}{2} \frac{\alpha \langle q^4 \rangle}{N^3} \right) (\delta_{c_{ij},2} - 2\delta_{c_{ij},1} + \delta_{c_{ij},0}) \\
 &\quad + \frac{\alpha^3 \langle q^2 \rangle^3}{6N^3} (\delta_{c_{ij},3} - 3\delta_{c_{ij},2} + 3\delta_{c_{ij},1} - \delta_{c_{ij},0}) + \mathcal{O}(N^{-4}) \tag{A.1}
 \end{aligned}$$

From this one reads off directly the values of  $p(c_{ij} = 0)$ ,  $p(c_{ij} = 1)$  and  $p(c_{ij} \geq 2)$ . The density of triangles is obtained writing (A.2) as

$$m_3 = (N-1)(N-2) \sum_{\mu\nu\rho=1}^{\alpha N} \langle \xi^{\mu} \xi^{\nu} \rangle \langle \xi^{\nu} \xi^{\rho} \rangle \langle \xi^{\rho} \xi^{\mu} \rangle \tag{A.2}$$

and using

$$\langle \xi^{\mu} \xi^{\nu} \rangle = \langle \xi^{\mu} \rangle \langle \xi^{\nu} \rangle + \delta_{\mu\nu} \langle \xi^{\mu} \rangle (1 - \langle \xi^{\mu} \rangle) = \frac{q_{\mu} q_{\nu}}{N^2} + \delta_{\mu\nu} \frac{q_{\mu}}{N} \left(1 - \frac{q_{\mu}}{N}\right) \tag{A.3}$$

This gives

$$\begin{aligned}
m_3 &= \frac{1}{N} \left[ 1 + \mathcal{O}\left(\frac{1}{N}\right) \right] \sum_{\mu\nu\rho=1}^{\alpha N} q_\mu q_\nu q_\rho \left[ \frac{q_\nu}{N} + \delta_{\mu\nu} \left(1 - \frac{q_\nu}{N}\right) \right] \left[ \frac{q_\rho}{N} + \delta_{\nu\rho} \left(1 - \frac{q_\rho}{N}\right) \right] \left[ \frac{q_\mu}{N} + \delta_{\rho\mu} \left(1 - \frac{q_\mu}{N}\right) \right] \\
&= \frac{1}{N} \left( 1 + \mathcal{O}\left(\frac{1}{N}\right) \right) \sum_{\mu\nu\rho=1}^{\alpha N} q_\mu q_\nu q_\rho \left\{ \frac{q_\mu q_\nu q_\rho}{N^3} + 3\delta_{\mu\nu} \frac{q_\rho q_\mu}{N^2} \left(1 - \frac{q_\mu}{N}\right) + 3\delta_{\mu\nu} \delta_{\nu\rho} \frac{q_\mu}{N} \left(1 - \frac{q_\mu}{N}\right)^2 \right. \\
&\quad \left. + \delta_{\mu\nu} \delta_{\nu\rho} \delta_{\rho\mu} \left(1 - \frac{q_\mu}{N}\right)^3 \right\} \\
&= \frac{1}{N} \sum_{\mu=1}^{\alpha N} \left(1 - \frac{q_\mu}{N}\right)^3 q_\mu^3 + \mathcal{O}\left(\frac{1}{N}\right) = \alpha \langle q^3 \rangle + \mathcal{O}(N^{-1})
\end{aligned} \tag{A.4}$$

### Appendix A.2. The $d$ -ensemble

In the  $d$ -ensemble we obtain

$$\begin{aligned}
p(c_{ij}) &= \langle \delta_{c_{ij}, \sum_\mu \xi_i^\mu \xi_j^\mu} \rangle = \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega c_{ij} + \frac{d_i d_j}{\alpha N} (e^{-i\omega} - 1) - \frac{1}{2} \frac{d_i^2 d_j^2}{(\alpha N)^3} (e^{-i\omega} - 1)^2} \\
&= \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega c_{ij}} \left[ 1 + \frac{d_i d_j}{\alpha N} (e^{-i\omega} - 1) + \frac{1}{2} \left( \frac{d_i d_j}{\alpha N} \right)^2 (e^{-i\omega} - 1)^2 - \frac{1}{2} \frac{(d_i d_j)^2}{(\alpha N)^3} (e^{-i\omega} - 1)^2 \right. \\
&\quad \left. + \frac{1}{6} \left( \frac{d_i d_j}{\alpha N} \right)^3 (e^{-i\omega} - 1)^3 + \dots \right]
\end{aligned} \tag{A.5}$$

which gives

$$\begin{aligned}
p(c_{ij} = 0) &= 1 - \frac{d_i d_j}{\alpha N} + \frac{1}{2} \left( \frac{d_i d_j}{\alpha N} \right)^2 - \frac{1}{6} \left( \frac{d_i d_j}{\alpha N} \right)^3 - \frac{1}{2} \frac{d_i^2 d_j^2}{(\alpha N)^3} \\
p(c_{ij} = 1) &= \frac{d_i d_j}{\alpha N} - \left( \frac{d_i d_j}{\alpha N} \right)^2 + \frac{1}{2} \left( \frac{d_i d_j}{\alpha N} \right)^3 + \frac{d_i^2 d_j^2}{(\alpha N)^3} \\
p(c_{ij} \geq 2) &= \mathcal{O}(N^{-2})
\end{aligned} \tag{A.6}$$

### Appendix A.3. The mixed ensemble

For the mixed ensemble, the link likelihood is found to be

$$\begin{aligned}
p(c_{ij}) &= \langle \delta_{c_{ij}, \sum_\mu \xi_i^\mu \xi_j^\mu} \rangle = \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega c_{ij} + \sum_\mu \frac{d_i d_j q_\mu^2}{\alpha^2 \langle q \rangle^2 N^2} (e^{-i\omega} - 1)} = \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega c_{ij} + \frac{d_i d_j \langle q^2 \rangle}{\alpha \langle q \rangle^2 N} (e^{-i\omega} - 1)} \\
&= e^{-\frac{d_i d_j \langle q^2 \rangle}{\alpha \langle q \rangle^2 N}} \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega c_{ij}} \left[ 1 + \frac{d_i d_j \langle q^2 \rangle}{\alpha N \langle q \rangle^2} e^{-i\omega} + \frac{1}{2} \left( \frac{d_i d_j \langle q^2 \rangle}{\alpha N \langle q \rangle^2} \right)^2 e^{-2i\omega} + \dots \right]
\end{aligned} \tag{A.7}$$

giving

$$\begin{aligned}
p(c_{ij} = 0) &= 1 - \frac{d_i d_j \langle q^2 \rangle}{\alpha \langle q \rangle^2 N} + \frac{1}{2} \left( \frac{d_i d_j \langle q^2 \rangle}{\alpha \langle q \rangle^2 N} \right)^2 + \mathcal{O}(N^{-3}) \\
p(c_{ij} = 1) &= \frac{d_i d_j \langle q^2 \rangle}{\alpha \langle q \rangle^2 N} \left( 1 - \frac{d_i d_j \langle q^2 \rangle}{\alpha \langle q \rangle^2 N} \right) + \mathcal{O}(N^{-3}) \\
p(c_{ij} \geq 2) &= \mathcal{O}(N^{-2})
\end{aligned} \tag{A.8}$$

## Appendix B. Calculation of the degree distribution $p(k)$

In this appendix we calculate the degree distribution of the weighted protein interaction network  $c_{ij} = \sum_{\mu} \xi_i^{\mu} \xi_j^{\mu}$ , in which the entries  $\xi_i^{\mu}$  are drawn from the bipartite ensembles (1,3,5), respectively.

### Appendix B.1. The $q$ -ensemble

In the  $q$ -ensemble, we can calculate  $p(k)$  as follows:

$$\begin{aligned}
p(k) &= \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k} \left\langle \frac{1}{N} \sum_i e^{-i\omega \sum_j c_{ij}} \right\rangle_{\xi} = \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k} \left\langle e^{-i\omega \sum_{j>1} c_{1j}} \right\rangle_{\xi} \\
&= \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k} \left\langle e^{-i\omega \sum_{\mu} \xi_1^{\mu} \sum_{j>1} \xi_j^{\mu}} \right\rangle_{\xi} = \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k} \prod_{\mu} \left\langle e^{-i\omega \xi_1^{\mu} \sum_{j>1} \xi_j^{\mu}} \right\rangle_{\xi} \\
&= \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k} \prod_{\mu} \left\{ 1 + \frac{q_{\mu}}{N} \left[ \left\langle e^{-i\omega \xi^{\mu}} \right\rangle_{\xi^{\mu}}^{N-1} - 1 \right] \right\} \\
&= \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k} \prod_{\mu} \left\{ 1 + \frac{q_{\mu}}{N} \left[ \left( 1 + \frac{q_{\mu}}{N} (e^{-i\omega} - 1) \right)^{N-1} - 1 \right] \right\} \\
&= \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k + \sum_{\mu} \frac{q_{\mu}}{N} \left[ \exp[q_{\mu}(e^{-i\omega} - 1)] - 1 \right] + \mathcal{O}(N^{-1})} \\
&= \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k + \alpha} \left\langle \exp[q(e^{-i\omega} - 1)] - 1 \right\rangle + \mathcal{O}(N^{-1}) \\
&= e^{-\alpha\langle q \rangle} \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k + \alpha} \left\langle q e^{-q} \exp[q e^{-i\omega}] \right\rangle + \mathcal{O}(N^{-1}) \\
&= e^{-\alpha\langle q \rangle} \sum_{\ell \geq 0} \frac{\alpha^{\ell}}{\ell!} \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k} \left\langle q e^{-q} \exp[q e^{-i\omega}] \right\rangle^{\ell} + \mathcal{O}(N^{-1}) \\
&= e^{-\alpha\langle q \rangle} \sum_{\ell \geq 0} \frac{\alpha^{\ell}}{\ell!} \left\langle \prod_{r \leq \ell} (q_r e^{-q_r}) \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k} e^{e^{-i\omega} \sum_{r \leq \ell} q_r} \right\rangle_{q_1 \dots q_{\ell}} + \mathcal{O}(N^{-1}) \\
&= e^{-\alpha\langle q \rangle} \sum_{\ell \geq 0} \frac{\alpha^{\ell}}{\ell!} \left\langle \prod_{r \leq \ell} (q_r e^{-q_r}) \sum_{s \geq 0} \frac{(\sum_{r \leq \ell} q_r)^s}{s!} \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k - i\omega s} \right\rangle_{q_1 \dots q_{\ell}} + \mathcal{O}(N^{-1}) \\
&= e^{-\alpha\langle q \rangle} \sum_{\ell \geq 0} \frac{\alpha^{\ell}}{\ell!} \left\langle \prod_{r \leq \ell} (q_r e^{-q_r}) \frac{(\sum_{r \leq \ell} q_r)^k}{k!} \right\rangle_{q_1 \dots q_{\ell}} + \mathcal{O}(N^{-1}) \tag{B.1}
\end{aligned}$$

Hence, for large network sizes  $N \rightarrow \infty$  we obtain

$$\begin{aligned}
\lim_{N \rightarrow \infty} p(k) &= e^{-\alpha\langle q \rangle} \sum_{\ell \geq 0} \frac{\alpha^{\ell}}{\ell!} \left\langle \left( \prod_{r \leq \ell} q_r \right) e^{-\sum_{r \leq \ell} q_r} \frac{(\sum_{r \leq \ell} q_r)^k}{k!} \right\rangle_{q_1 \dots q_{\ell}} \\
&= e^{-\alpha\langle q \rangle} \sum_{\ell \geq 0} \frac{\alpha^{\ell}}{\ell!} \sum_{q_1 \dots q_{\ell} \geq 0} p(q_1) \dots p(q_{\ell}) q_1 \dots q_{\ell} e^{-\sum_{r \leq \ell} q_r} \frac{(\sum_{r \leq \ell} q_r)^k}{k!} \tag{B.2}
\end{aligned}$$

We can rewrite this in terms of the distribution  $W(q) = qP(q)/\langle q \rangle$ , which denotes the likelihood to draw a link attached to a node of degree  $q$  in the bi-partite graph,

$$\lim_{N \rightarrow \infty} p(k) = e^{-\alpha\langle q \rangle} \sum_{\ell \geq 0} \frac{(\alpha\langle q \rangle)^{\ell}}{\ell!} \sum_{q_1 \dots q_{\ell} \geq 0} W(q_1) \dots W(q_{\ell}) e^{-\sum_{r \leq \ell} q_r} \frac{(\sum_{r \leq \ell} q_r)^k}{k!} \tag{B.3}$$

and upon defining

$$P(y) = e^{-\alpha\langle q \rangle} \sum_{\ell \geq 0} \frac{(\alpha\langle q \rangle)^\ell}{\ell!} \sum_{q_1 \dots q_\ell \geq 0} W(q_1) \dots W(q_\ell) \delta[y - \sum_{r \leq \ell} q_r] \quad (\text{B.4})$$

we finally get to

$$\lim_{N \rightarrow \infty} p(k) = \int_0^\infty dy P(y) e^{-y} y^k / k! \quad (\text{B.5})$$

The interpretation is that if we draw  $\ell$  from a Poisson distribution with  $\langle \ell \rangle = \alpha\langle q \rangle$ , and then draw  $\ell$  variables  $q_r$  from  $W(q_r)$ , we find  $k$  as a Poissonian variable with  $\langle k \rangle = \sum_{r \leq \ell} q_r$ . Clearly  $p(k)$  is normalised, and for its first moment we find:

$$\begin{aligned} \langle k \rangle &= \int_0^\infty dy P(y) y = e^{-\alpha\langle q \rangle} \sum_{\ell \geq 0} \frac{(\alpha\langle q \rangle)^\ell}{\ell!} \sum_{q_1 \dots q_\ell \geq 0} W(q_1) \dots W(q_\ell) \sum_{r \leq \ell} q_r \\ &= e^{-\alpha\langle q \rangle} \sum_{\ell > 0} \frac{(\alpha\langle q \rangle)^\ell}{(\ell-1)!} \sum_q W(q) q = \alpha\langle q^2 \rangle \end{aligned} \quad (\text{B.6})$$

For the second moment we obtain

$$\begin{aligned} \langle k^2 \rangle &= \langle k \rangle + \int_0^\infty dy P(y) y^2 \\ &= \alpha\langle q^2 \rangle + e^{-\alpha\langle q \rangle} \sum_{\ell \geq 0} \frac{(\alpha\langle q \rangle)^\ell}{\ell!} \sum_{q_1 \dots q_\ell \geq 0} W(q_1) \dots W(q_\ell) \sum_{r,s \leq \ell} q_r q_s \\ &= \alpha\langle q^2 \rangle + e^{-\alpha\langle q \rangle} \left( \sum_q W(q) q \right)^2 \sum_{\ell > 0} \frac{(\alpha\langle q \rangle)^\ell}{\ell!} \ell^2 \\ &\quad + e^{-\alpha\langle q \rangle} \left[ \sum_q W(q) q^2 - \left( \sum_q W(q) q \right)^2 \right] \sum_{\ell > 0} \frac{(\alpha\langle q \rangle)^\ell}{(\ell-1)!} \\ &= \alpha\langle q^2 \rangle + e^{-\alpha\langle q \rangle} \sum_{\ell \geq 0} \frac{(\alpha\langle q \rangle)^\ell}{\ell!} \sum_{q_1 \dots q_\ell \geq 0} W(q_1) \dots W(q_\ell) \sum_{r,s \leq \ell} q_r q_s \\ &= \alpha\langle q^2 \rangle + \alpha \left[ \langle q^3 \rangle - \frac{\langle q^2 \rangle^2}{\langle q \rangle} \right] + \frac{\langle q^2 \rangle^2}{\langle q \rangle^2} e^{-\alpha\langle q \rangle} \sum_{\ell > 0} \frac{(\alpha\langle q \rangle)^\ell}{\ell!} \ell^2 \\ &= \alpha\langle q^2 \rangle + \alpha \left[ \langle q^3 \rangle - \frac{\langle q^2 \rangle^2}{\langle q \rangle} \right] + \frac{\langle q^2 \rangle^2}{\langle q \rangle^2} [\alpha^2 \langle q \rangle^2 + \alpha\langle q \rangle] \\ &= \alpha\langle q^2 \rangle + \alpha\langle q^3 \rangle + \alpha^2 \langle q^2 \rangle^2 \end{aligned} \quad (\text{B.7})$$

This is in agreement with results from a direct calculation:

$$\begin{aligned} \langle k^2 \rangle &= \frac{1}{N} \sum_{i \neq j \neq k} \langle c_{ij} c_{kl} \rangle = \frac{1}{N} \sum_{i \neq j} \langle c_{ij} c_{ji} \rangle + \frac{1}{N} \sum_{[ijk]} \langle c_{ij} c_{jk} \rangle \\ &= \frac{1}{N} \sum_{i \neq j} \sum_{\mu\nu} \langle \xi_i^\mu \xi_j^\mu \xi_i^\nu \xi_j^\nu \rangle + \frac{1}{N} \sum_{[ijk]} \sum_{\mu\nu} \langle \xi_i^\mu \xi_j^\mu \xi_j^\nu \xi_k^\nu \rangle \\ &= \frac{1}{N} \sum_{i \neq j} \sum_{\mu} \langle \xi_i^\mu \xi_j^\mu \rangle + \frac{1}{N} \sum_{[ijk]} \sum_{\mu \neq \nu} \langle \xi_i^\mu \xi_j^\mu \rangle \langle \xi_j^\nu \xi_k^\nu \rangle + \frac{1}{N} \sum_{[ijk]} \sum_{\mu} \langle \xi_i^\mu \xi_j^\mu \xi_k^\mu \rangle + \mathcal{O}(N^{-1}) \\ &= \frac{1}{N} \sum_{i \neq j} \sum_{\mu} \frac{q_\mu^2}{N^2} + \frac{1}{N} \sum_{[ijk]} \sum_{\mu \neq \nu} \frac{q_\mu^2}{N^2} \frac{q_\nu^2}{N^2} + \frac{1}{N} \sum_{[ijk]} \sum_{\mu} \frac{q_\mu^3}{N^3} + \mathcal{O}(N^{-1}) \\ &= \alpha\langle q^2 \rangle + (\alpha\langle q^2 \rangle)^2 + \alpha\langle q^3 \rangle + \mathcal{O}(N^{-1}) = \langle k \rangle + \langle k \rangle^2 + \alpha\langle q^3 \rangle + \mathcal{O}(N^{-1}) \end{aligned} \quad (\text{B.8})$$

Appendix B.2. The  $d$ -ensemble

We can calculate the asymptotic degree distribution in the  $d$ -ensemble as follows

$$\begin{aligned}
p(k) &= \lim_{N \rightarrow \infty} \frac{1}{N} \sum_i \langle \delta_{k, \sum_j c_{ij}} \rangle_{\xi} = \frac{1}{N} \sum_i \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k} \langle e^{-i\omega \sum_{\mu} \xi_i^{\mu} \sum_j \xi_j^{\mu}} \rangle_{\xi} \\
&= \lim_{N \rightarrow \infty} \frac{1}{N} \sum_i \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k} \prod_{\mu} \left[ 1 + \frac{d_i}{\alpha N} \left( \prod_j \langle e^{-i\omega \xi_j^{\mu}} \rangle - 1 \right) \right] \\
&= \lim_{N \rightarrow \infty} \frac{1}{N} \sum_i \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k} \prod_{\mu} \left[ 1 + \frac{d_i}{\alpha N} \left( e^{\frac{\langle d \rangle}{\alpha} (e^{-i\omega} - 1)} - 1 \right) \right] \\
&= \lim_{N \rightarrow \infty} \frac{1}{N} \sum_i \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k + d_i \left( e^{\frac{\langle d \rangle}{\alpha} (e^{-i\omega} - 1)} - 1 \right)} \\
&= \sum_d P(d) \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k + d \left( e^{\frac{\langle d \rangle}{\alpha} (e^{-i\omega} - 1)} - 1 \right)} \\
&= \sum_d P(d) e^{-d} \sum_{\ell} \frac{d^{\ell}}{\ell!} e^{-\ell \frac{\langle d \rangle}{\alpha}} \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k + \ell \frac{\langle d \rangle}{\alpha} e^{-i\omega}} \\
&= \sum_d P(d) \sum_{\ell} e^{-d} \frac{d^{\ell}}{\ell!} e^{-\ell \frac{\langle d \rangle}{\alpha}} \frac{\left( \frac{\ell \langle d \rangle}{\alpha} \right)^k}{k!}
\end{aligned} \tag{B.9}$$

## Appendix B.3. The mixed ensemble

In the mixed ensemble we have the asymptotic degree distribution

$$\begin{aligned}
p(k) &= \lim_{N \rightarrow \infty} \frac{1}{N} \sum_i \langle \delta_{k, \sum_j c_{ij}} \rangle_{\xi} = \frac{1}{N} \sum_i \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k} \langle e^{-i\omega \sum_{\mu} \xi_i^{\mu} \sum_j \xi_j^{\mu}} \rangle_{\xi} \\
&= \lim_{N \rightarrow \infty} \frac{1}{N} \sum_i \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k} \prod_{\mu} \left[ 1 + \frac{d_i q_{\mu}}{\alpha \langle q \rangle N} \left( \prod_j \langle e^{-i\omega \xi_j^{\mu}} \rangle - 1 \right) \right] \\
&= \lim_{N \rightarrow \infty} \frac{1}{N} \sum_i \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k} \prod_{\mu} \left[ 1 + \frac{d_i q_{\mu}}{\alpha \langle q \rangle N} \left( e^{q_{\mu} (e^{-i\omega} - 1)} - 1 \right) \right] \\
&= \lim_{N \rightarrow \infty} \frac{1}{N} \sum_i \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k + \sum_{\mu} \frac{d_i q_{\mu}}{\alpha \langle q \rangle N} \left( e^{q_{\mu} (e^{-i\omega} - 1)} - 1 \right)} \\
&= \sum_d P(d) \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k + \frac{d}{\langle q \rangle} \langle q (e^{q(e^{-i\omega} - 1)} - 1) \rangle_q} \\
&= \sum_d P(d) e^{-d} \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k + \frac{d}{\langle q \rangle} \langle q e^{-q} \exp[q e^{-i\omega}] \rangle_q} \\
&= \sum_d P(d) e^{-d} \sum_{\ell \geq 0} \frac{(d/\langle q \rangle)^{\ell}}{\ell!} \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k} \langle q e^{-q} \exp[q e^{-i\omega}] \rangle_q^{\ell} \\
&= \sum_d P(d) e^{-d} \sum_{\ell \geq 0} \frac{d^{\ell}}{\ell!} \left\langle \prod_{r \leq \ell} \left( \frac{q_r e^{-q_r}}{\langle q \rangle} \right) \frac{(\sum_{r \leq \ell} q_r)^k}{k!} \right\rangle_{q_1 \dots q_{\ell}}
\end{aligned} \tag{B.10}$$

We can rewrite this expression in terms of the associated distribution  $W(q) = qP(q)/\langle q \rangle$  as:

$$p(k) = \sum_d P(d) e^{-d} \sum_{\ell \geq 0} \frac{d^{\ell}}{\ell!} \left\langle \prod_{r \leq \ell} \left( \frac{q_r e^{-q_r}}{\langle q \rangle} \right) \frac{(\sum_{r \leq \ell} q_r)^k}{k!} \right\rangle_{q_1 \dots q_{\ell}}$$



$$= \sum_d P(d) e^{-d} \sum_{\ell \geq 0} \frac{d^\ell}{\ell!} \sum_{q_1 \dots q_\ell \geq 0} W(q_1) \dots W(q_\ell) e^{-\sum_{r \leq \ell} q_r} \frac{(\sum_{r \leq \ell} q_r)^k}{k!} \quad (\text{B.11})$$

or, equivalently, as

$$p(k) = \int_0^\infty dy P(y) e^{-y} y^k / k! \quad (\text{B.12})$$

where

$$P(y) = \sum_d P(d) e^{-d} \sum_{\ell \geq 0} \frac{d^\ell}{\ell!} \sum_{q_1 \dots q_\ell \geq 0} W(q_1) \dots W(q_\ell) \delta[y - \sum_{r \leq \ell} q_r] \quad (\text{B.13})$$

The first two moments of  $p(k)$  are

$$\begin{aligned} \langle k \rangle &= \int_0^\infty dy P(y) y = \sum_d P(d) e^{-d} \sum_{\ell \geq 0} \frac{d^\ell}{\ell!} \sum_{q_1 \dots q_\ell \geq 0} W(q_1) \dots W(q_\ell) \sum_{r \leq \ell} q_r \\ &= \sum_d P(d) e^{-d} \sum_{\ell > 0} \frac{d^\ell}{(\ell-1)!} \sum_q W(q) q = \langle d \rangle \frac{\langle q^2 \rangle}{\langle q \rangle} = \alpha \langle q^2 \rangle \end{aligned} \quad (\text{B.14})$$

$$\begin{aligned} \langle k^2 \rangle &= \langle k \rangle + \int_0^\infty dy P(y) y^2 \\ &= \alpha \langle q^2 \rangle + \sum_d P(d) e^{-d} \sum_{\ell \geq 0} \frac{d^\ell}{\ell!} \sum_{q_1 \dots q_\ell \geq 0} W(q_1) \dots W(q_\ell) \sum_{r,s \leq \ell} q_r q_s \\ &= \alpha \langle q^2 \rangle + \sum_d P(d) e^{-d} \sum_{\ell > 0} \frac{d^\ell}{\ell!} \left[ \ell \sum_q W(q) q^2 + \ell(\ell-1) \left( \sum_q W(q) q \right)^2 \right] \\ &= \alpha \langle q^2 \rangle + \frac{\langle q^3 \rangle}{\langle q \rangle} \langle d \rangle + \langle d^2 \rangle \frac{\langle q^2 \rangle^2}{\langle q \rangle^2} = \alpha \langle q^2 \rangle + \alpha \langle q^3 \rangle + \frac{\langle d^2 \rangle}{\langle d \rangle^2} \langle k \rangle^2 \end{aligned} \quad (\text{B.15})$$

## Appendix C. The link between observables in the $\mathbf{a}$ and $\mathbf{c}$ networks

In this appendix we inspect the relation between expectation values of various observables in the ensembles  $p(\mathbf{a})$  and  $p(\mathbf{c})$ .

### Appendix C.1. The $q$ -ensemble

Denoting averages in the  $\mathbf{a}$  ensemble as  $\langle \dots \rangle_a$ , we have, for the  $q$ -ensemble of bipartite graphs:

$$\begin{aligned} \langle k \rangle_a &= \frac{1}{N} \sum_{ij} \langle a_{ij} \rangle_a = \frac{1}{N} \sum_{ij} \langle \theta[c_{ij} - \frac{1}{2}] \rangle \\ &= \frac{1}{N} \sum_{ij} [1 - \langle \delta_{c_{ij},0} \rangle] = \alpha \langle q^2 \rangle + \mathcal{O}(N^{-1}) = \langle k \rangle + \mathcal{O}(N^{-1}) \\ \langle k^2 \rangle_a &= \frac{1}{N} \sum_{i \neq j \neq k} \langle a_{ij} a_{jk} \rangle = \frac{1}{N} \sum_{ij} \langle a_{ij} \rangle + \frac{1}{N} \sum_{i \neq j \neq k (\neq i)} \langle a_{ij} a_{jk} \rangle \\ &= \frac{1}{N} \sum_{ij} \langle (1 - \delta_{c_{ij},0}) \rangle + \frac{1}{N} \sum_{i \neq j \neq k (\neq i)} \langle (1 - \delta_{c_{ij},0}) (1 - \delta_{c_{jk},0}) \rangle \\ &= \frac{1}{N} \sum_{ij} \frac{\alpha \langle q^2 \rangle}{N} + \frac{1}{N} \sum_{i \neq j \neq k (\neq i)} (1 - 2 \langle \delta_{c_{ij},0} \rangle + \langle \delta_{c_{ij},0} \delta_{c_{jk},0} \rangle) \end{aligned} \quad (\text{C.1})$$

$$\begin{aligned}
&= (N-1)(N-2) - 2(N-1)(N-2) \left(1 - \frac{\alpha \langle q^2 \rangle}{N} + \frac{\alpha^2 \langle q^2 \rangle^2}{2N^2}\right) \\
&\quad + (N-1)(N-2) \left(1 - 2\frac{\alpha \langle q^2 \rangle}{N} + \frac{\alpha \langle q^3 \rangle}{N^2} + 2\frac{\alpha^2 \langle q^2 \rangle^2}{N^2}\right) + \alpha \langle q^2 \rangle \\
&= \alpha \langle q^2 \rangle + \alpha \langle q^3 \rangle + \alpha^2 \langle q^2 \rangle^2 \equiv \langle k^2 \rangle
\end{aligned} \tag{C.2}$$

where we used

$$\begin{aligned}
&\frac{1}{N(N-1)(N-2)} \sum_{i \neq j \neq k (\neq i)} \langle \delta_{c_{ij},0} \delta_{c_{jk},0} \rangle \\
&= \frac{1}{N(N-1)(N-2)} \sum_{i \neq j \neq k (\neq i)} \int_{-\pi}^{\pi} \frac{d\omega d\omega'}{4\pi^2} \prod_{\mu} \langle e^{i\xi_j^{\mu}(\xi_i^{\mu}\omega + \xi_k^{\mu}\omega')} \rangle \\
&= \int_{-\pi}^{\pi} \frac{d\omega d\omega'}{4\pi^2} \prod_{\mu} \left\{ 1 + \frac{q_{\mu}^2}{N^2} \left[ (e^{i\omega} + e^{i\omega'} - 2) + \frac{q_{\mu}}{N} (e^{i(\omega+\omega')} - e^{i\omega} - e^{i\omega'} + 1) \right] \right\} \\
&= \int_{-\pi}^{\pi} \frac{d\omega d\omega'}{4\pi^2} e^{\frac{\alpha \langle q^2 \rangle}{N} (e^{i\omega} + e^{i\omega'} - 2) + \frac{\alpha \langle q^3 \rangle}{N^2} (e^{i(\omega+\omega')} - e^{i\omega} - e^{i\omega'} + 1) - \frac{\alpha \langle q^4 \rangle}{2N^3} (e^{i\omega} + e^{i\omega'} - 2)^2} \\
&= 1 - 2\frac{\alpha \langle q^2 \rangle}{N} + \frac{\alpha \langle q^3 \rangle}{N^2} + 2\frac{\alpha^2 \langle q^2 \rangle^2}{N^2} - 2\frac{\alpha \langle q^4 \rangle}{N^3} - 2\frac{\alpha^2 \langle q^2 \rangle \langle q^3 \rangle}{N^3} - \frac{4}{3} \frac{\alpha^3 \langle q^2 \rangle^3}{N^3}
\end{aligned} \tag{C.3}$$

For loops of length 3 we proceed in the same way, obtaining

$$\begin{aligned}
m_3^a &= \frac{1}{N} \sum_{i \neq j \neq k (\neq i)} \langle a_{ij} a_{jk} a_{ki} \rangle = \frac{1}{N} \sum_{i \neq j \neq k (\neq i)} \langle (1 - \delta_{c_{ij},0})(1 - \delta_{c_{jk},0})(1 - \delta_{c_{ki},0}) \rangle \\
&= \frac{1}{N} \sum_{i \neq j \neq k (\neq i)} (1 - 3\langle \delta_{c_{ij},0} \rangle + 3\langle \delta_{c_{ij},0} \delta_{c_{jk},0} \rangle - \langle \delta_{c_{ij},0} \delta_{c_{jk},0} \delta_{c_{ki},0} \rangle) \\
&= (N-1)(N-2) - 3(N-1)(N-2) \left(1 - \frac{\alpha \langle q^2 \rangle}{N} + 2\frac{\alpha^2 \langle q^2 \rangle^2}{2N^2}\right) \\
&\quad + 3(N-1)(N-2) \left(1 - 2\frac{\alpha \langle q^2 \rangle}{N} + \frac{\alpha \langle q^3 \rangle}{N^2} + 2\frac{\alpha^2 \langle q^2 \rangle^2}{N^2}\right) \\
&\quad - (N-1)(N-2) \left(1 - 3\frac{\alpha \langle q^2 \rangle}{N} + 2\frac{\alpha \langle q^3 \rangle}{N^2} + \frac{9}{2} \frac{\alpha^2 \langle q^2 \rangle^2}{N^2}\right) = \alpha \langle q^3 \rangle \equiv m_3^c
\end{aligned} \tag{C.4}$$

where we used

$$\begin{aligned}
&\frac{1}{N(N-1)(N-2)} \sum_{i \neq j \neq k (\neq i)} \langle \delta_{c_{ij},0} \delta_{c_{jk},0} \delta_{c_{ki},0} \rangle \\
&= \frac{1}{N(N-1)(N-2)} \sum_{i \neq j \neq k (\neq i)} \int_{-\pi}^{\pi} \frac{d\omega d\omega' d\omega''}{8\pi^3} \prod_{\mu} \langle e^{i\xi_i^{\mu}(\xi_j^{\mu}\omega + \xi_k^{\mu}\omega' + \xi_l^{\mu}\omega'')} \rangle \\
&= \int_{-\pi}^{\pi} \frac{d\omega d\omega' d\omega''}{8\pi^3} \prod_{\mu} \left\{ 1 + \frac{q_{\mu}^2}{N^2} \left[ (e^{i\omega} + e^{i\omega'} + e^{i\omega''} - 3) + \frac{q_{\mu}}{N} (e^{i(\omega+\omega'+\omega'')} - e^{i\omega} - e^{i\omega'} - e^{i\omega''} + 2) \right] \right\} \\
&= \int_{-\pi}^{\pi} \frac{d\omega d\omega' d\omega''}{8\pi^3} e^{\sum_{\mu} \frac{q_{\mu}^2}{N^2} (e^{i\omega} + e^{i\omega'} + e^{i\omega''} - 3) + \sum_{\mu} \frac{q_{\mu}^3}{N^3} (e^{i(\omega+\omega'+\omega'')} - e^{i\omega} - e^{i\omega'} - e^{i\omega''} + 2) - \sum_{\mu} \frac{q_{\mu}^4}{2N^4} (e^{i\omega} + e^{i\omega'} + e^{i\omega''} - 3)^2} \\
&= \int_{-\pi}^{\pi} \frac{d\omega d\omega' d\omega''}{8\pi^3} e^{\frac{\alpha \langle q^2 \rangle}{N} (e^{i\omega} + e^{i\omega'} + e^{i\omega''} - 3) + \frac{\alpha \langle q^3 \rangle}{N^2} (e^{i(\omega+\omega'+\omega'')} - e^{i\omega} - e^{i\omega'} - e^{i\omega''} + 2) - \frac{\alpha \langle q^4 \rangle}{2N^3} (e^{i\omega} + e^{i\omega'} + e^{i\omega''} - 3)^2} \\
&= 1 - 3\frac{\alpha \langle q^2 \rangle}{N} + 2\frac{\alpha \langle q^3 \rangle}{N^2} + \frac{9}{2} \frac{\alpha^2 \langle q^2 \rangle^2}{N^2} - \frac{9}{2} \frac{\alpha \langle q^4 \rangle}{N^3} - 6\frac{\alpha^2 \langle q^2 \rangle \langle q^3 \rangle}{N^3} - \frac{9}{2} \frac{\alpha^3 \langle q^2 \rangle^3}{N^3}
\end{aligned} \tag{C.5}$$

Finally for loops of length 4, we have

$$\begin{aligned}
m_4^a &= \frac{1}{N} \sum_{[i,j,k,\ell]} \langle a_{ij} a_{jk} a_{k\ell} a_{\ell i} \rangle \\
&= \frac{1}{N} \sum_{[i,j,k,\ell]} \langle (1 - \delta_{c_{ij},0})(1 - \delta_{c_{jk},0})(1 - \delta_{c_{k\ell},0})(1 - \delta_{c_{\ell i},0}) \rangle \\
&= \frac{1}{N} \sum_{[i,j,k,\ell]} (1 - 4\langle \delta_{c_{ij},0} \rangle + 4\langle \delta_{c_{ij},0} \delta_{c_{jk},0} \rangle + 2\langle \delta_{c_{ij},0} \rangle \langle \delta_{c_{jk},0} \rangle - 4\langle \delta_{c_{ij},0} \delta_{c_{jk},0} \delta_{c_{k\ell},0} \rangle \\
&\quad + \langle \delta_{c_{ij},0} \delta_{c_{jk},0} \delta_{c_{k\ell},0} \delta_{c_{\ell i},0} \rangle) \\
&= (N-1)(N-2)(N-3) \left\{ 1 - 4 \left( 1 - \frac{\alpha \langle q^2 \rangle}{N} + \frac{\alpha^2 \langle q^2 \rangle^2}{2N^2} - \frac{\alpha^3 \langle q^2 \rangle^3}{6N^3} - \frac{\alpha \langle q^4 \rangle}{2N^3} \right) \right. \\
&\quad + 4 \left( 1 - 2 \frac{\alpha \langle q^2 \rangle}{N} + \frac{\alpha \langle q^3 \rangle}{N^2} + 2 \frac{\alpha^2 \langle q^2 \rangle^2}{N^2} - \frac{4}{3} \frac{\alpha^3 \langle q^2 \rangle^3}{N^3} - 2 \frac{\alpha \langle q^4 \rangle}{N^3} - 2 \frac{\alpha^2 \langle q^2 \rangle \langle q^3 \rangle}{N^3} \right) \\
&\quad + 2 \left( 1 - \frac{\alpha \langle q^2 \rangle}{N} + \frac{\alpha^2 \langle q^2 \rangle^2}{2N^2} - \frac{\alpha^3 \langle q^2 \rangle^3}{6N^3} - \frac{\alpha \langle q^4 \rangle}{2N^3} \right)^2 \\
&\quad - 4 \left( 1 - 3 \frac{\alpha \langle q^2 \rangle}{N} + 2 \frac{\alpha \langle q^3 \rangle}{N^2} + \frac{9}{2} \frac{\alpha^2 \langle q^2 \rangle^2}{N^2} - \frac{9}{2} \frac{\alpha^3 \langle q^2 \rangle^3}{N^3} - \frac{9}{2} \frac{\alpha \langle q^4 \rangle}{N^3} - 6 \frac{\alpha^2 \langle q^2 \rangle \langle q^3 \rangle}{N^3} \right) \\
&\quad \left. + \left( 1 - 4 \frac{\alpha \langle q^2 \rangle}{N} + 4 \frac{\alpha \langle q^3 \rangle}{N^2} - 9 \frac{\alpha}{N^3} \langle q^4 \rangle + 8 \frac{\alpha^2 \langle q^2 \rangle^2}{N^2} - 16 \frac{\alpha^2 \langle q^2 \rangle \langle q^3 \rangle}{N^3} - \frac{32}{3} \frac{\alpha^3 \langle q^2 \rangle^3}{N^3} \right) \right\} \\
&= \alpha \langle q^4 \rangle \equiv m_4^c \tag{C.6}
\end{aligned}$$

where we used

$$\begin{aligned}
&\frac{1}{N(N-1)(N-2)(N-3)} \sum_{i \neq j \neq k (\neq i)} \langle \delta_{c_{ij},0} \delta_{c_{jk},0} \delta_{c_{k\ell},0} \rangle \\
&= \frac{1}{N(N-1)(N-2)(N-3)} \sum_{i \neq j \neq k (\neq i)} \int_{-\pi}^{\pi} \frac{d\omega d\omega' d\omega''}{8\pi^3} \prod_{\mu} \langle e^{i\xi_j^{\mu} (\xi_i^{\mu} \omega + \xi_k^{\mu} \omega') + i\xi_{\ell}^{\mu} \xi_k^{\mu} \omega''} \rangle \\
&= \int_{-\pi}^{\pi} \frac{d\omega d\omega' d\omega''}{8\pi^3} \prod_{\mu} \left\{ 1 + \frac{q_{\mu}^2}{N^2} \left[ (e^{i\omega} + e^{i\omega'} + e^{i\omega''} - 3) + \frac{q_{\mu}}{N} (e^{i\omega'} - 1)(e^{i\omega} + e^{i\omega''} - 2) \right. \right. \\
&\quad \left. \left. + \frac{q_{\mu}^2}{N^2} e^{i\omega'} (e^{i\omega} - 1)(e^{i\omega''} - 1) \right] \right\} \\
&= \int_{-\pi}^{\pi} \frac{d\omega d\omega' d\omega''}{8\pi^3} e^{\sum_{\mu} \frac{q_{\mu}^2}{N^2} (e^{i\omega} + e^{i\omega'} + e^{i\omega''} - 3) + \sum_{\mu} \frac{q_{\mu}^3}{N^3} (e^{i\omega'} - 1)(e^{i\omega} + e^{i\omega''} - 2) - \sum_{\mu} \frac{q_{\mu}^4}{2N^4} (e^{i\omega} + e^{i\omega'} + e^{i\omega''} - 3)^2} \\
&\quad \times e^{\frac{q_{\mu}^4}{N^4} e^{i\omega'} (e^{i\omega} - 1)(e^{i\omega''} - 1)} \\
&= \int_{-\pi}^{\pi} \frac{d\omega d\omega' d\omega''}{8\pi^3} e^{\frac{\alpha \langle q^2 \rangle}{N} (e^{i\omega} + e^{i\omega'} + e^{i\omega''} - 3) + \frac{\alpha \langle q^3 \rangle}{N^2} (e^{i\omega'} - 1)(e^{i\omega} + e^{i\omega''} - 2) - \frac{\alpha \langle q^4 \rangle}{2N^3} (e^{i\omega} + e^{i\omega'} + e^{i\omega''} - 3)^2} \\
&\quad \times e^{\alpha \frac{\langle q^4 \rangle}{N^3} e^{i\omega'} (e^{i\omega} - 1)(e^{i\omega''} - 1)} \\
&= 1 - 3 \frac{\alpha \langle q^2 \rangle}{N} + 2 \frac{\alpha \langle q^3 \rangle}{N^2} + \frac{9}{2} \frac{\alpha^2 \langle q^2 \rangle^2}{N^2} - \frac{9}{2} \frac{\alpha \langle q^4 \rangle}{N^3} - 6 \frac{\alpha^2 \langle q^2 \rangle \langle q^3 \rangle}{N^3} - \frac{9}{2} \frac{\alpha^3 \langle q^2 \rangle^3}{N^3} \tag{C.7}
\end{aligned}$$

and

$$\frac{1}{N(N-1)(N-2)(N-3)} \sum_{[ijk\ell]} \langle \delta_{c_{ij},0} \delta_{c_{jk},0} \delta_{c_{k\ell},0} \delta_{c_{\ell i},0} \rangle$$

$$\begin{aligned}
&= \frac{1}{N(N-1)(N-2)(N-3)} \sum_{[ijkl]} \int_{-\pi}^{\pi} \frac{d\omega d\omega' d\omega'' d\omega'''}{16\pi^4} \prod_{\mu} \langle e^{i\xi_i^{\mu}(\xi_j^{\mu}\omega + \xi_{\ell}^{\mu}\omega''') + \xi_k^{\mu}(\xi_j^{\mu}\omega' + \xi_{\ell}^{\mu}\omega'')} \rangle \\
&= \int_{-\pi}^{\pi} \frac{d\omega d\omega' d\omega'' d\omega'''}{16\pi^4} \prod_{\mu} \left\{ \left(1 - \frac{q_{\mu}}{N}\right) \left\{ \frac{q_{\mu}}{N} \left[ \frac{q_{\mu}^2}{N^2} e^{i(\omega+\omega')} + \frac{q_{\mu}}{N} \left(1 - \frac{q_{\mu}}{N}\right) (e^{i\omega'} + e^{i\omega''}) + \left(1 - \frac{q_{\mu}}{N}\right)^2 \right] \right. \right. \\
&\quad \left. \left. + \left(1 - \frac{q_{\mu}}{N}\right) \right\} + \frac{q_{\mu}}{N} \left\{ \frac{q_{\mu}^2}{N^2} e^{i(\omega+\omega''')} \left(1 - \frac{q_{\mu}}{N} + \frac{q_{\mu}}{N} e^{i(\omega'+\omega'')}\right) \right. \right. \\
&\quad \left. \left. + \frac{q_{\mu}}{N} \left(1 - \frac{q_{\mu}}{N}\right) \left[ e^{i\omega} \left(1 - \frac{q_{\mu}}{N} + \frac{q_{\mu}}{N} e^{i\omega'}\right) + e^{i\omega'''} \left(1 - \frac{q_{\mu}}{N} + \frac{q_{\mu}}{N} e^{i\omega''}\right) \right] + \left(1 - \frac{q_{\mu}}{N}\right)^2 \right\} \right\} \\
&= \prod_{\mu} \left\{ \frac{q_{\mu}}{N} \left(1 - \frac{q_{\mu}}{N}\right)^2 + \left(1 - \frac{q_{\mu}}{N}\right) \left[ \frac{q_{\mu}}{N} \left(1 - \frac{q_{\mu}}{N}\right)^2 + \left(1 - \frac{q_{\mu}}{N}\right) \right] \right\} \\
&= \prod_{\mu} \left\{ 1 - 4 \frac{q_{\mu}^2}{N^2} + 4 \frac{q_{\mu}^3}{N^3} - \frac{q_{\mu}^4}{N^4} \right\} = e^{-4 \frac{\alpha \langle q^2 \rangle}{N} + 4 \frac{\alpha \langle q^3 \rangle}{N^2} - 9 \frac{\alpha \langle q^4 \rangle}{N^3} + \mathcal{O}(N^{-4})} \\
&= 1 - 4 \frac{\alpha \langle q^2 \rangle}{N} + 4 \frac{\alpha \langle q^3 \rangle}{N^2} + 8 \frac{\alpha^2 \langle q^2 \rangle^2}{N^2} - 9 \frac{\alpha \langle q^4 \rangle}{N^3} - 16 \frac{\alpha^2 \langle q^2 \rangle \langle q^3 \rangle}{N^3} - \frac{32}{3} \frac{\alpha^3 \langle q^2 \rangle^3}{N^3} + \mathcal{O}(N^{-4}) \quad (C.8)
\end{aligned}$$

Again, the square brackets underneath the summations indicate that all indices are different, to exclude backtracking in the counting of loops of length 4.

### Appendix C.2. The $d$ -ensemble

For the  $d$ -ensemble, denoting averages relating to  $\mathbf{a}$  as  $\langle \dots \rangle_a$ , we have:

$$\begin{aligned}
\langle k \rangle_a &= \frac{1}{N} \sum_{ij} \langle a_{ij} \rangle_a = \frac{1}{N} \sum_{ij} [1 - \langle \delta_{c_{ij},0} \rangle] = \\
&= \frac{1}{N} \sum_{ij} \left[ \frac{d_i d_j}{\alpha N} - \frac{1}{2} \left( \frac{d_i d_j}{\alpha N} \right)^2 + \frac{1}{6} \left( \frac{d_i d_j}{\alpha N} \right)^3 + \frac{1}{2} \frac{d_i^2 d_j^2}{(\alpha N)^3} \right] \\
&= \frac{\langle d \rangle^2}{\alpha} + \mathcal{O}(N^{-1}) = \langle k \rangle + \mathcal{O}(N^{-1}) \quad (C.9) \\
\langle k^2 \rangle_a &= \frac{1}{N} \sum_{i \neq j \neq k} \langle a_{ij} a_{jk} \rangle = \frac{1}{N} \sum_{ij} \langle a_{ij} \rangle + \frac{1}{N} \sum_{i \neq j \neq k (\neq i)} \langle a_{ij} a_{jk} \rangle \\
&= \frac{\langle d \rangle^2}{\alpha} + \frac{1}{N} \sum_{i \neq j \neq k (\neq i)} \langle (1 - \delta_{c_{ij},0}) (1 - \delta_{c_{jk},0}) \rangle \\
&= \frac{\langle d \rangle^2}{\alpha} + \frac{1}{N} \sum_{i \neq j \neq k (\neq i)} (1 - 2 \langle \delta_{c_{ij},0} \rangle + \langle \delta_{c_{ij},0} \delta_{c_{jk},0} \rangle) \\
&= \frac{\langle d \rangle^2}{\alpha} + (N-1)(N-2) - \frac{2}{N} \sum_{[ijk]} \left( 1 - \frac{d_i d_j}{\alpha N} + \frac{1}{2} \left( \frac{d_i d_j}{\alpha N} \right)^2 \right) + \frac{1}{N} \sum_{[ijk]} \langle \delta_{c_{ij},0} \delta_{c_{jk},0} \rangle \\
&= \frac{\langle d \rangle^2}{\alpha} + 2 \frac{\langle d \rangle^2}{\alpha} N - 2 \frac{\langle d \rangle^2}{\alpha} - \frac{\langle d^2 \rangle^2}{\alpha^2} - 2N \frac{\langle d \rangle^2}{\alpha} + 2 \frac{\langle d \rangle^2}{\alpha} + \frac{\langle d \rangle^3}{\alpha^2} + \frac{\langle d^2 \rangle^2}{\alpha^2} - \frac{\langle d^2 \rangle \langle d \rangle^2}{\alpha^2} \\
&= \frac{\langle d \rangle^2}{\alpha} + \frac{\langle d \rangle^3}{\alpha^2} + \frac{\langle d^2 \rangle \langle d \rangle^2}{\alpha^2} \equiv \langle k^2 \rangle \quad (C.10)
\end{aligned}$$

where we used

$$\begin{aligned}
\frac{1}{N} \sum_{i \neq j \neq k (\neq i)} \langle \delta_{c_{ij},0} \delta_{c_{jk},0} \rangle &= \frac{1}{N} \sum_{i \neq j \neq k (\neq i)} \int_{-\pi}^{\pi} \frac{d\omega d\omega'}{4\pi^2} \prod_{\mu} \langle e^{i\xi_j^{\mu}(\xi_i^{\mu}\omega + \xi_k^{\mu}\omega')} \rangle \\
&= \frac{1}{N} \sum_{[ijk]} \int_{-\pi}^{\pi} \frac{d\omega d\omega'}{4\pi^2} \prod_{\mu} \left\{ 1 + \frac{d_j}{\alpha N} \left[ \frac{d_i}{\alpha N} (e^{i\omega} - 1) + \frac{d_k}{\alpha N} (e^{i\omega'} - 1) + \frac{d_i d_k}{(\alpha N)^2} (e^{i(\omega+\omega')} - e^{i\omega} - e^{i\omega'} + 1) \right] \right\} \\
&= \frac{1}{N} \sum_{[ijk]} \int_{-\pi}^{\pi} \frac{d\omega d\omega'}{4\pi^2} \left\{ 1 + \frac{d_j}{\alpha N} \left[ d_i (e^{i\omega} - 1) + d_k (e^{i\omega'} - 1) + \frac{d_i d_k}{\alpha N} (e^{i(\omega+\omega')} - e^{i\omega} - e^{i\omega'} + 1) \right] \right. \\
&\quad \left. + \frac{1}{2} \left( \frac{d_j}{\alpha N} [d_i (e^{i\omega} - 1) + d_k (e^{i\omega'} - 1)] \right)^2 - \frac{d_i d_j^2 d_k}{(\alpha N)^3} (d_i + d_k) \right\} \\
&= \frac{1}{N} \sum_{[ijk]} \left\{ 1 + \frac{d_j}{\alpha N} \left[ -d_i - d_k + \frac{d_i d_k}{\alpha N} \right] + \frac{1}{2} \left( \frac{d_j}{\alpha N} \right)^2 (d_i^2 + d_k^2 + 2d_i d_k) - \frac{d_i^2 d_j^2 d_k}{(\alpha N)^3} - \frac{d_i d_j^2 d_k^2}{(\alpha N)^3} \right\} \\
&= (N-1)(N-2) - 2N \frac{\langle d \rangle^2}{\alpha} + 2 \frac{\langle d \rangle^2}{\alpha} + \frac{\langle d \rangle^3}{\alpha^2} + \frac{\langle d^2 \rangle^2}{\alpha^2} + \frac{\langle d^2 \rangle \langle d \rangle^2}{\alpha^2} \tag{C.11}
\end{aligned}$$

For loops of length 3 we have:

$$\begin{aligned}
m_3^a &= \frac{1}{N} \sum_{i \neq j \neq k (\neq i)} \langle a_{ij} a_{jk} a_{ki} \rangle = \frac{1}{N} \sum_{i \neq j \neq k (\neq i)} \langle (1 - \delta_{c_{ij},0})(1 - \delta_{c_{jk},0})(1 - \delta_{c_{ki},0}) \rangle \\
&= \frac{1}{N} \sum_{[ijk]} (1 - 3\langle \delta_{c_{ij},0} \rangle + 3\langle \delta_{c_{ij},0} \delta_{c_{jk},0} \rangle - \langle \delta_{c_{ij},0} \delta_{c_{jk},0} \delta_{c_{ki},0} \rangle) \\
&= (N-1)(N-2) - 3 \frac{1}{N} \sum_{[ijk]} \left( 1 - \frac{d_i d_j}{\alpha N} + \frac{1}{2} \left( \frac{d_i d_j}{\alpha N} \right)^2 \right) \\
&\quad + 3 \left[ (N-1)(N-2) - 2N \frac{\langle d \rangle^2}{\alpha} + 2 \frac{\langle d \rangle^2}{\alpha} + \frac{\langle d \rangle^3}{\alpha^2} + \frac{\langle d^2 \rangle^2}{\alpha^2} - \frac{\langle d^2 \rangle \langle d \rangle^2}{\alpha^2} \right] \\
&\quad - (N-1)(N-2) + 3N \frac{\langle d \rangle^2}{\alpha} - 3 \frac{\langle d \rangle^2}{\alpha} - 2 \frac{\langle d \rangle^3}{\alpha^2} - \frac{3}{2} \frac{\langle d^2 \rangle^2}{\alpha^2} - 3 \frac{\langle d^2 \rangle \langle d \rangle^2}{\alpha^2} \\
&= 3 \frac{\langle d \rangle^2}{\alpha} N - 3 \frac{\langle d \rangle^2}{\alpha} - \frac{3}{2} \frac{\langle d^2 \rangle^2}{\alpha^2} \\
&\quad + 3 \left[ -2N \frac{\langle d \rangle^2}{\alpha} + 2 \frac{\langle d \rangle^2}{\alpha} + \frac{\langle d \rangle^3}{\alpha^2} + \frac{\langle d^2 \rangle^2}{\alpha^2} + \frac{\langle d^2 \rangle \langle d \rangle^2}{\alpha^2} \right] \\
&\quad + 3N \frac{\langle d \rangle^2}{\alpha} - 3 \frac{\langle d \rangle^2}{\alpha} - 2 \frac{\langle d \rangle^3}{\alpha^2} - \frac{3}{2} \frac{\langle d^2 \rangle^2}{\alpha^2} - 3 \frac{\langle d^2 \rangle \langle d \rangle^2}{\alpha^2} = \frac{\langle d \rangle^3}{\alpha^2} \equiv m_3^c \tag{C.12}
\end{aligned}$$

where we used

$$\begin{aligned}
\frac{1}{N} \sum_{[ijk]} \langle \delta_{c_{ij},0} \delta_{c_{jk},0} \delta_{c_{ki},0} \rangle &= \frac{1}{N} \sum_{[ijk]} \int_{-\pi}^{\pi} \frac{d\omega d\omega' d\omega''}{8\pi^3} \prod_{\mu} \langle e^{i\xi_i^{\mu}(\xi_j^{\mu}\omega + \xi_k^{\mu}\omega'') + i\xi_j^{\mu}\xi_k^{\mu}\omega'} \rangle \\
&= \frac{1}{N} \sum_{[ijk]} \int_{-\pi}^{\pi} \frac{d\omega d\omega' d\omega''}{8\pi^3} \prod_{\mu} \left\{ \frac{d_i}{\alpha N} \langle e^{i(\xi_j^{\mu}\omega + \xi_k^{\mu}\omega'' + \xi_j^{\mu}\xi_k^{\mu}\omega')} \rangle + \left( 1 - \frac{d_i}{\alpha N} \right) \langle e^{i\xi_j^{\mu}\xi_k^{\mu}\omega'} \rangle \right\} \\
&= \frac{1}{N} \sum_{[ijk]} \int_{-\pi}^{\pi} \frac{d\omega d\omega' d\omega''}{8\pi^3} \left( 1 + \frac{d_j d_k}{(\alpha N)^2} (e^{i\omega'} - 1) + \frac{d_i}{\alpha N} \left\{ -1 - \frac{d_j d_k}{(\alpha N)^2} (e^{i\omega'} - 1) \right\} \right)
\end{aligned}$$

$$\begin{aligned}
& + \frac{d_j}{\alpha N} e^{i\omega} \left[ 1 + \frac{d_k}{\alpha N} (e^{i(\omega'' + \omega')} - 1) \right] + \left( 1 - \frac{d_j}{\alpha N} \right) \left[ 1 + \frac{d_k}{\alpha N} (e^{i\omega''} - 1) \right] \Bigg\}^{\alpha N} \\
& = \frac{1}{N} \sum_{[ijk]} \left( 1 - \frac{d_j d_k}{(\alpha N)^2} + \frac{d_i}{\alpha N} \left\{ -\frac{d_j}{\alpha N} - \frac{d_k}{\alpha N} + 2 \frac{d_j d_k}{(\alpha N)^2} \right\} \right)^{\alpha N} \\
& = \frac{1}{N} \sum_{[ijk]} \left[ 1 - \frac{d_j d_k}{\alpha N} + \frac{d_i}{\alpha N} \left\{ -d_j - d_k + 2 \frac{d_j d_k}{\alpha N} \right\} + \frac{1}{2} \left( \frac{d_i}{\alpha N} \right)^2 (d_j^2 + d_k^2 + 2d_j d_k) \right. \\
& \quad \left. + \frac{1}{2} \frac{d_j^2 d_k^2}{(\alpha N)^2} + \frac{d_i d_j d_k}{(\alpha N)^2} (d_j + d_k) \right] \\
& = (N-1)(N-2) - 3N \frac{\langle d \rangle^2}{\alpha} + 3 \frac{\langle d \rangle^2}{\alpha} + 2 \frac{\langle d \rangle^3}{\alpha^2} + \frac{3}{2} \frac{\langle d^2 \rangle^2}{\alpha^2} + 3 \frac{\langle d^2 \rangle \langle d \rangle^2}{\alpha^2} \tag{C.13}
\end{aligned}$$

Finally, for loops of length 4 we have

$$\begin{aligned}
m_4^a & = \frac{1}{N} \sum_{[i,j,k,\ell]} \langle a_{ij} a_{jk} a_{k\ell} a_{\ell i} \rangle = \frac{1}{N} \sum_{[i,j,k,\ell]} \langle (1 - \delta_{c_{ij},0})(1 - \delta_{c_{jk},0})(1 - \delta_{c_{k\ell},0})(1 - \delta_{c_{\ell i},0}) \rangle \\
& = \frac{1}{N} \sum_{[i,j,k,\ell]} (1 - 4\langle \delta_{c_{ij},0} \rangle + 4\langle \delta_{c_{ij},0} \delta_{c_{jk},0} \rangle + 2\langle \delta_{c_{ij},0} \rangle \langle \delta_{c_{k\ell},0} \rangle - 4\langle \delta_{c_{ij},0} \delta_{c_{jk},0} \delta_{c_{k\ell},0} \rangle) \\
& \quad + \langle \delta_{c_{ij},0} \delta_{c_{jk},0} \delta_{c_{k\ell},0} \delta_{c_{\ell i},0} \rangle \\
& = (N-1)(N-2)(N-3) - 4 \left[ (N-1)(N-2)(N-3) - N^2 \frac{\langle d \rangle^2}{\alpha} \right. \\
& \quad \left. + \frac{1}{2} N \frac{\langle d^2 \rangle^2}{\alpha^2} - \frac{1}{2} \frac{\langle d^2 \rangle^2}{\alpha^3} - \frac{1}{6} \frac{\langle d^3 \rangle^2}{\alpha^3} \right] \\
& + 4 \left[ (N-1)(N-2)(N-3) - 2N^2 \frac{\langle d \rangle^2}{\alpha} + N \frac{\langle d \rangle^3}{\alpha^2} - \frac{\langle d^2 \rangle^2}{\alpha^3} \right. \\
& \quad \left. - \frac{\langle d^2 \rangle \langle d \rangle^2}{\alpha^3} + N \frac{\langle d^2 \rangle^2}{\alpha^2} + N \frac{\langle d^2 \rangle \langle d \rangle^2}{\alpha^2} - 2 \frac{\langle d^2 \rangle^2 \langle d \rangle}{\alpha^3} - \frac{1}{3} \frac{\langle d^3 \rangle^2}{\alpha^3} - \frac{\langle d^3 \rangle \langle d^2 \rangle \langle d \rangle}{\alpha^3} \right] \\
& + 2 \left[ (N-1)(N-2)(N-3) - 2N^2 \frac{\langle d \rangle^2}{\alpha} + N \frac{\langle d^2 \rangle^2}{\alpha^2} + N \frac{\langle d \rangle^4}{\alpha^2} \right. \\
& \quad \left. - \frac{\langle d^2 \rangle^2}{\alpha^3} - \frac{1}{3} \frac{\langle d^3 \rangle^2}{\alpha^3} - \frac{\langle d^2 \rangle^2 \langle d \rangle^2}{\alpha^3} \right] \\
& - 4 \left[ (N-1)(N-2)(N-3) - 3N^2 \frac{\langle d \rangle^2}{\alpha} + 2N \frac{\langle d \rangle^3}{\alpha^2} - \frac{3}{2} \frac{\langle d^2 \rangle^2}{\alpha^3} \right. \\
& \quad \left. - 2 \frac{\langle d^2 \rangle \langle d \rangle^2}{\alpha^3} + \frac{3}{2} N \frac{\langle d^2 \rangle^2}{\alpha^2} - \frac{\langle d \rangle^4}{\alpha^3} + 2N \frac{\langle d^2 \rangle \langle d \rangle^2}{\alpha^2} + N \frac{\langle d \rangle^4}{\alpha^2} \right. \\
& \quad \left. - 4 \frac{\langle d^2 \rangle^2 \langle d \rangle}{\alpha^3} - 2 \frac{\langle d^2 \rangle \langle d \rangle^3}{\alpha^3} - \frac{1}{2} \frac{\langle d^3 \rangle^2}{\alpha^3} - 2 \frac{\langle d^3 \rangle \langle d^2 \rangle \langle d \rangle}{\alpha^3} - \frac{\langle d^2 \rangle^2 \langle d \rangle^2}{\alpha^3} \right] \\
& + (N-1)(N-2)(N-3) - 4N^2 \frac{\langle d \rangle^2}{\alpha} + 4N \frac{\langle d \rangle^3}{\alpha^2} - 2 \frac{\langle d^2 \rangle^2}{\alpha^3} \\
& - 4 \frac{\langle d^2 \rangle \langle d \rangle^2}{\alpha^3} + 2N \frac{\langle d^2 \rangle^2}{\alpha^2} - 3 \frac{\langle d \rangle^4}{\alpha^3} + 4N \frac{\langle d^2 \rangle \langle d \rangle^2}{\alpha^2} + 2N \frac{\langle d \rangle^4}{\alpha^2} \\
& - 8 \frac{\langle d^2 \rangle^2 \langle d \rangle}{\alpha^3} - 8 \frac{\langle d^2 \rangle \langle d \rangle^3}{\alpha^3} - \frac{2}{3} \frac{\langle d^3 \rangle^2}{\alpha^3} - 4 \frac{\langle d^3 \rangle \langle d^2 \rangle \langle d \rangle}{\alpha^3} - 2 \frac{\langle d^2 \rangle^2 \langle d \rangle^2}{\alpha^3}
\end{aligned}$$

$$= \frac{\langle d \rangle^4}{\alpha^3} \equiv m_4^c \quad (\text{C.14})$$

where we used

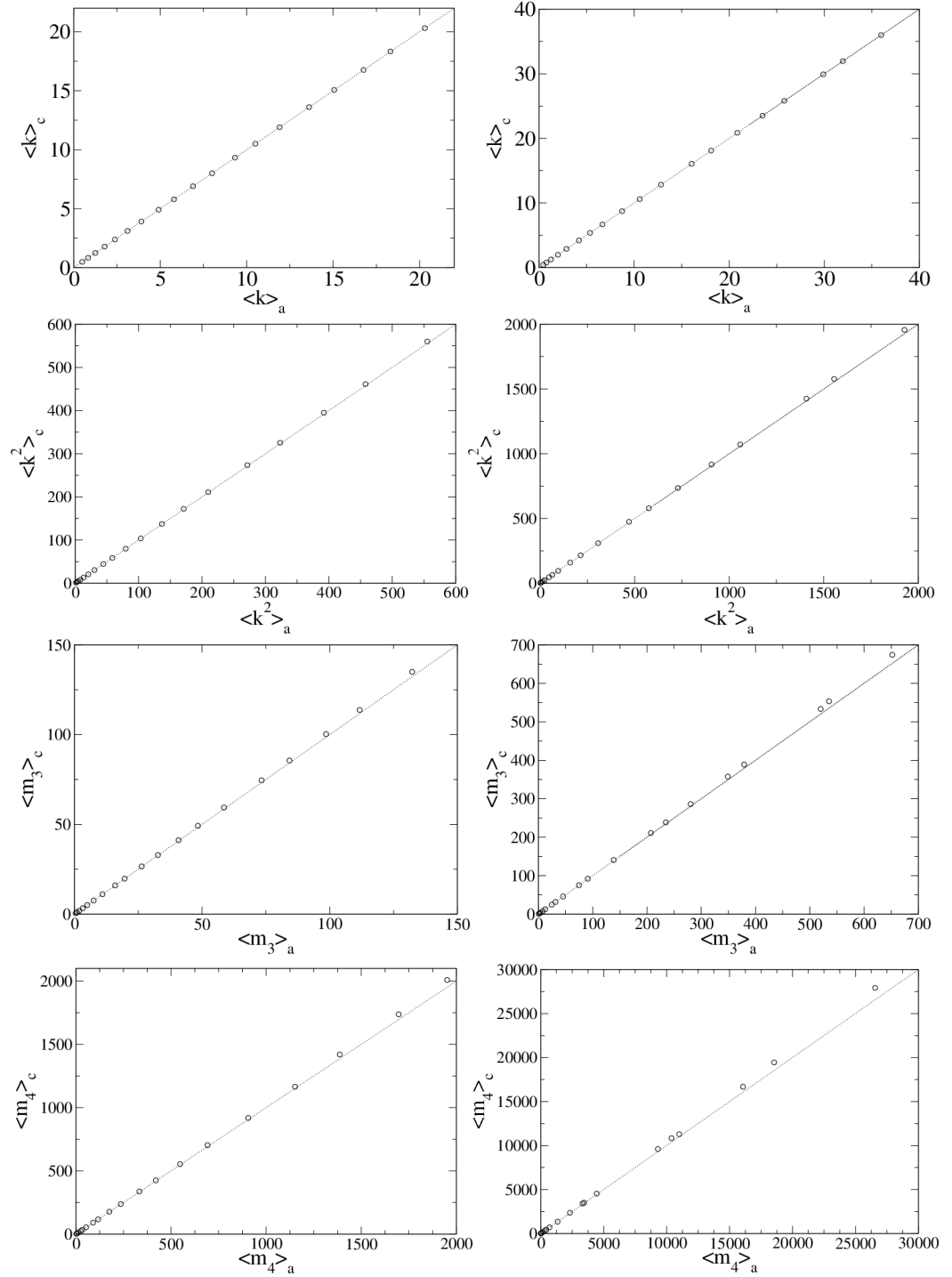
$$\begin{aligned}
\frac{1}{N} \sum_{[ijkl]} \langle \delta_{c_{ij},0} \delta_{c_{jk},0} \delta_{c_{kl},0} \rangle &= \frac{1}{N} \sum_{[ijkl]} \int_{-\pi}^{\pi} \frac{d\omega d\omega' d\omega''}{8\pi^3} \prod_{\mu} \langle e^{i\xi_j^{\mu}(\xi_i^{\mu}\omega + \xi_k^{\mu}\omega') + \xi_k^{\mu}\xi_{\ell}^{\mu}\omega''} \rangle \\
&= \frac{1}{N} \sum_{[ijkl]} \int_{-\pi}^{\pi} \frac{d\omega d\omega' d\omega''}{8\pi^3} \prod_{\mu} \left\{ \frac{d_j}{\alpha N} \langle e^{i(\xi_i^{\mu}\omega + \xi_k^{\mu}\omega') + i\xi_k^{\mu}\xi_{\ell}^{\mu}\omega''} \rangle + \left(1 - \frac{d_j}{\alpha N}\right) \langle e^{i\xi_k^{\mu}\xi_{\ell}^{\mu}\omega''} \rangle \right\} \\
&= \frac{1}{N} \sum_{[ijkl]} \int_{-\pi}^{\pi} \frac{d\omega d\omega' d\omega''}{8\pi^3} \left( 1 + \frac{d_i d_j}{(\alpha N)^2} (e^{i\omega} - 1) + \frac{d_j d_k}{(\alpha N)^2} (e^{i\omega'} - 1) + \frac{d_k d_{\ell}}{(\alpha N)^2} (e^{i\omega''} - 1) \right. \\
&\quad + \frac{d_i d_j d_k}{(\alpha N)^3} (e^{i(\omega+\omega')} - e^{i\omega} - e^{i\omega'} + 1) + \frac{d_j d_k d_{\ell}}{(\alpha N)^3} (e^{i(\omega'+\omega'')} - e^{i\omega'} - e^{i\omega''} + 1) \\
&\quad \left. + \frac{d_i d_j d_k d_{\ell}}{(\alpha N)^4} (e^{i(\omega+\omega'+\omega'')} - e^{i(\omega+\omega')} - e^{i(\omega+\omega'')} + e^{i\omega'}) \right)^{\alpha N} \\
&= \frac{1}{N} \sum_{[ijkl]} \left\{ 1 - \frac{d_i d_j}{\alpha N} - \frac{d_j d_k}{\alpha N} - \frac{d_k d_{\ell}}{\alpha N} + \frac{d_i d_j d_k}{(\alpha N)^2} + \frac{d_j d_k d_{\ell}}{(\alpha N)^2} \right. \\
&\quad - \frac{1}{2} \left[ \frac{d_i^2 d_j^2}{(\alpha N)^3} + \frac{d_j^2 d_k^2}{(\alpha N)^3} + \frac{d_k^2 d_{\ell}^2}{(\alpha N)^3} + 2 \frac{d_i d_j^2 d_k}{(\alpha N)^3} + 2 \frac{d_i d_j d_k d_{\ell}}{(\alpha N)^3} + 2 \frac{d_j d_k^2 d_{\ell}}{(\alpha N)^3} \right] \\
&\quad + \frac{1}{2} \left[ \frac{d_i^2 d_j^2}{(\alpha N)^2} + \frac{d_j^2 d_k^2}{(\alpha N)^2} + \frac{d_k^2 d_{\ell}^2}{(\alpha N)^2} + 2 \frac{d_i d_j^2 d_k}{(\alpha N)^2} + 2 \frac{d_i d_j d_k d_{\ell}}{(\alpha N)^2} + 2 \frac{d_j d_k^2 d_{\ell}}{(\alpha N)^2} - 2 \frac{d_i^2 d_j^2 d_k}{(\alpha N)^3} \right. \\
&\quad - 2 \frac{d_i^2 d_j d_k d_{\ell}}{(\alpha N)^3} - 2 \frac{d_i d_j^2 d_k^2}{(\alpha N)^3} - 2 \frac{d_i d_j d_k^2 d_{\ell}}{(\alpha N)^3} - 2 \frac{d_j^2 d_k^2 d_{\ell}}{(\alpha N)^3} - 2 \frac{d_j d_k^2 d_{\ell}^2}{(\alpha N)^3} \left. \right] - \frac{1}{6} \left[ \frac{d_i^3 d_j^3}{(\alpha N)^3} + \frac{d_j^3 d_k^3}{(\alpha N)^3} \right. \\
&\quad \left. + \frac{d_k^3 d_{\ell}^3}{(\alpha N)^3} + 3 \frac{d_i^2 d_j^3 d_k}{(\alpha N)^3} + 3 \frac{d_i d_j^2 d_k^2}{(\alpha N)^3} + 3 \frac{d_i^2 d_j^2 d_k d_{\ell}}{(\alpha N)^3} + 3 \frac{d_i d_j d_k^2 d_{\ell}^2}{(\alpha N)^3} + 3 \frac{d_j^2 d_k^3 d_{\ell}}{(\alpha N)^3} + 3 \frac{d_j d_k^3 d_{\ell}^2}{(\alpha N)^3} \right] \Big\} \\
&= (N-1)(N-2)(N-3) - 3N^2 \frac{\langle d \rangle^2}{\alpha} + 2N \frac{\langle d \rangle^3}{\alpha^2} - \frac{3}{2} \frac{\langle d^2 \rangle^2}{\alpha^3} - 2 \frac{\langle d^2 \rangle \langle d \rangle^2}{\alpha^3} - \frac{\langle d \rangle^4}{\alpha^3} \\
&\quad - \frac{3}{2} N \frac{\langle d^2 \rangle^2}{\alpha^2} + 2N \frac{\langle d^2 \rangle \langle d \rangle^2}{\alpha^2} + N \frac{\langle d \rangle^4}{\alpha^2} - 4 \frac{\langle d^2 \rangle^2 \langle d \rangle}{\alpha^3} - 2 \frac{\langle d^2 \rangle \langle d \rangle^3}{\alpha^3} \\
&\quad - \frac{1}{2} \frac{\langle d^3 \rangle^2}{\alpha^3} - 2 \frac{\langle d^2 \rangle \langle d^2 \rangle \langle d \rangle}{\alpha^3} - \frac{\langle d^2 \rangle^2 \langle d \rangle^2}{\alpha^3} \quad (\text{C.15})
\end{aligned}$$

and

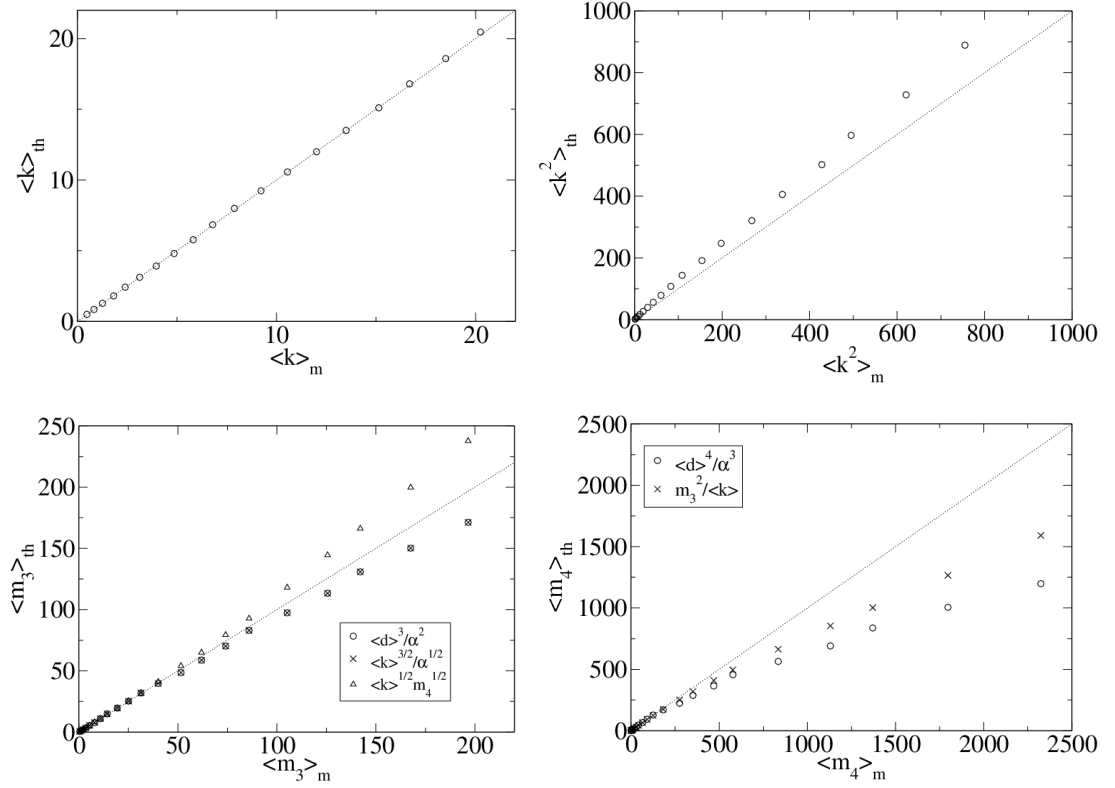
$$\begin{aligned}
\frac{1}{N} \sum_{[ijkl]} \langle \delta_{c_{ij},0} \delta_{c_{jk},0} \delta_{c_{kl},0} \delta_{c_{li},0} \rangle &= \frac{1}{N} \sum_{[ijkl]} \int_{-\pi}^{\pi} \frac{d\omega d\omega' d\omega'' d\omega'''}{16\pi^4} \prod_{\mu} \langle e^{i\xi_i^{\mu}(\xi_j^{\mu}\omega + \xi_{\ell}^{\mu}\omega''') + \xi_k^{\mu}(\xi_j^{\mu}\omega' + \xi_{\ell}^{\mu}\omega'')} \rangle \\
&= \frac{1}{N} \sum_{[ijkl]} \int_{-\pi}^{\pi} \frac{d\omega d\omega' d\omega'' d\omega'''}{16\pi^4} \prod_{\mu} \left\{ \frac{d_i}{\alpha N} \langle e^{i(\xi_j^{\mu}\omega + \xi_{\ell}^{\mu}\omega''') + i\xi_k^{\mu}(\xi_j^{\mu}\omega' + \xi_{\ell}^{\mu}\omega'')} \rangle + \left(1 - \frac{d_i}{\alpha N}\right) \langle e^{i\xi_k^{\mu}(\xi_j^{\mu}\omega' + \xi_{\ell}^{\mu}\omega'')} \rangle \right\} \\
&= \frac{1}{N} \sum_{[ijkl]} \int_{-\pi}^{\pi} \frac{d\omega d\omega' d\omega'' d\omega'''}{16\pi^4} \left( 1 + \frac{d_i d_j}{(\alpha N)^2} (e^{i\omega} - 1) + \frac{d_j d_k}{(\alpha N)^2} (e^{i\omega'} - 1) + \frac{d_k d_{\ell}}{(\alpha N)^2} (e^{i\omega''} - 1) \right. \\
&\quad + \frac{d_i d_{\ell}}{(\alpha N)^2} (e^{i\omega'''} - 1) + \frac{d_i d_j d_k}{(\alpha N)^3} (e^{i(\omega+\omega')} - e^{i\omega} - e^{i\omega'} + 1) + \frac{d_i d_k d_{\ell}}{(\alpha N)^3} (e^{i(\omega''+\omega''')} - e^{i\omega''} - e^{i\omega'''} + 1)
\end{aligned}$$

$$\begin{aligned}
& + \frac{d_i d_j d_\ell}{(\alpha N)^3} (e^{i(\omega+\omega''')} - e^{i\omega} - e^{i\omega'''} + 1) + \frac{d_j d_k d_\ell}{(\alpha N)^3} (e^{i(\omega'+\omega'')} - e^{i\omega'} - e^{i\omega''} + 1) \\
& + \frac{d_i d_j d_k d_\ell}{(\alpha N)^4} (e^{i(\omega+\omega'+\omega''+\omega''')} - e^{i(\omega+\omega')} - e^{i(\omega''+\omega''')} - e^{i(\omega+\omega''')} - e^{i(\omega'+\omega'')} \\
& + e^{i\omega} + e^{i\omega'} + e^{i\omega''} + e^{i\omega'''} - 1) \Big)^{\alpha N} \\
= & \frac{1}{N} \sum_{[ijkl]} \left\{ 1 - \frac{d_i d_j}{\alpha N} - \frac{d_j d_k}{\alpha N} - \frac{d_k d_\ell}{\alpha N} - \frac{d_i d_\ell}{\alpha N} + \frac{d_i d_j d_k}{(\alpha N)^2} + \frac{d_i d_k d_\ell}{(\alpha N)^2} + \frac{d_i d_j d_\ell}{(\alpha N)^2} + \frac{d_j d_k d_\ell}{(\alpha N)^2} \right. \\
& - \frac{d_i d_j d_k d_\ell}{(\alpha N)^3} - \frac{1}{2} \left[ \frac{d_i^2 d_j^2}{(\alpha N)^3} + \frac{d_j^2 d_k^2}{(\alpha N)^3} + \frac{d_k^2 d_\ell^2}{(\alpha N)^3} + \frac{d_i^2 d_\ell^2}{(\alpha N)^3} + 2 \frac{d_i d_j^2 d_k}{(\alpha N)^3} + 2 \frac{d_i d_j d_k d_\ell}{(\alpha N)^3} + 2 \frac{d_i^2 d_j d_\ell}{(\alpha N)^3} \right. \\
& + 2 \frac{d_j d_k^2 d_\ell}{(\alpha N)^3} + 2 \frac{d_i d_j d_k d_\ell}{(\alpha N)^3} + 2 \frac{d_i d_k d_\ell^2}{(\alpha N)^3} \Big] + \frac{1}{2} \left[ \frac{d_i^2 d_j^2}{(\alpha N)^2} + \frac{d_j^2 d_k^2}{(\alpha N)^2} + \frac{d_k^2 d_\ell^2}{(\alpha N)^2} + \frac{d_i^2 d_\ell^2}{(\alpha N)^2} + 2 \frac{d_i d_j^2 d_k}{(\alpha N)^2} \right. \\
& + 2 \frac{d_i d_j d_k d_\ell}{(\alpha N)^2} + 2 \frac{d_i^2 d_j d_\ell}{(\alpha N)^2} + 2 \frac{d_j d_k^2 d_\ell}{(\alpha N)^2} + 2 \frac{d_i d_j d_k d_\ell}{(\alpha N)^2} + 2 \frac{d_i d_k d_\ell^2}{(\alpha N)^2} - 2 \frac{d_i^2 d_j^2 d_k}{(\alpha N)^3} - 2 \frac{d_i^2 d_j d_\ell}{(\alpha N)^3} \\
& - 2 \frac{d_i^2 d_j d_k d_\ell}{(\alpha N)^3} - 2 \frac{d_i d_j^2 d_k d_\ell}{(\alpha N)^3} - 2 \frac{d_i d_j d_k^2 d_\ell}{(\alpha N)^3} - 2 \frac{d_i d_j d_k d_\ell^2}{(\alpha N)^3} - 2 \frac{d_i^2 d_j^2 d_k d_\ell}{(\alpha N)^3} - 2 \frac{d_i^2 d_j d_k d_\ell}{(\alpha N)^3} - 2 \frac{d_i d_j^2 d_k d_\ell}{(\alpha N)^3} \\
& - 2 \frac{d_i d_j d_k^2 d_\ell}{(\alpha N)^3} - 2 \frac{d_i d_j d_k d_\ell^2}{(\alpha N)^3} - 2 \frac{d_j d_k^2 d_\ell^2}{(\alpha N)^3} - 2 \frac{d_i^2 d_j d_k d_\ell^2}{(\alpha N)^3} - 2 \frac{d_i^2 d_k d_\ell^2}{(\alpha N)^3} - 2 \frac{d_i^2 d_j d_\ell^2}{(\alpha N)^3} - 2 \frac{d_i d_j d_k d_\ell^2}{(\alpha N)^3} \Big] \\
& - \frac{1}{6} \left[ \frac{d_i^3 d_j^3}{(\alpha N)^3} + \frac{d_j^3 d_k^3}{(\alpha N)^3} + \frac{d_k^3 d_\ell^3}{(\alpha N)^3} + \frac{d_i^3 d_\ell^3}{(\alpha N)^3} + 3 \frac{d_i^2 d_j^3 d_k}{(\alpha N)^3} + 3 \frac{d_i d_j^3 d_k^2}{(\alpha N)^3} + 3 \frac{d_i^2 d_j^2 d_k d_\ell}{(\alpha N)^3} \right. \\
& + 3 \frac{d_i d_j d_k^2 d_\ell^2}{(\alpha N)^3} + 3 \frac{d_i^3 d_j^2 d_\ell}{(\alpha N)^3} + 3 \frac{d_i^3 d_j d_\ell^2}{(\alpha N)^3} + 3 \frac{d_j^2 d_k^3 d_\ell}{(\alpha N)^3} + 3 \frac{d_j d_k^3 d_\ell^2}{(\alpha N)^3} + 3 \frac{d_i d_j^2 d_k^2 d_\ell}{(\alpha N)^3} + 3 \frac{d_i^2 d_j d_k d_\ell^2}{(\alpha N)^3} \\
& \left. + 3 \frac{d_i d_k^2 d_\ell^3}{(\alpha N)^3} + 3 \frac{d_i^2 d_k d_\ell^3}{(\alpha N)^3} \right] \Big\} \\
= & (N-1)(N-2)(N-3) - 4N^2 \frac{\langle d \rangle^2}{\alpha} + 4N \frac{\langle d \rangle^3}{\alpha^2} - 2 \frac{\langle d^2 \rangle^2}{\alpha^3} - \frac{\langle d \rangle^4}{\alpha^3} \\
& - 4 \frac{\langle d^2 \rangle \langle d \rangle^2}{\alpha^3} + 2N \frac{\langle d^2 \rangle^2}{\alpha^2} - 2 \frac{\langle d \rangle^4}{\alpha^3} + 4N \frac{\langle d^2 \rangle \langle d \rangle^2}{\alpha^2} + 2N \frac{\langle d \rangle^4}{\alpha^2} \\
& - 8 \frac{\langle d^2 \rangle^2 \langle d \rangle}{\alpha^3} - 8 \frac{\langle d^2 \rangle \langle d \rangle^3}{\alpha^3} - \frac{2 \langle d^3 \rangle^2}{3 \alpha^3} - 4 \frac{\langle d^3 \rangle \langle d^2 \rangle \langle d \rangle}{\alpha^3} - 2 \frac{\langle d^2 \rangle^2 \langle d \rangle^2}{\alpha^3}
\end{aligned} \tag{C.16}$$

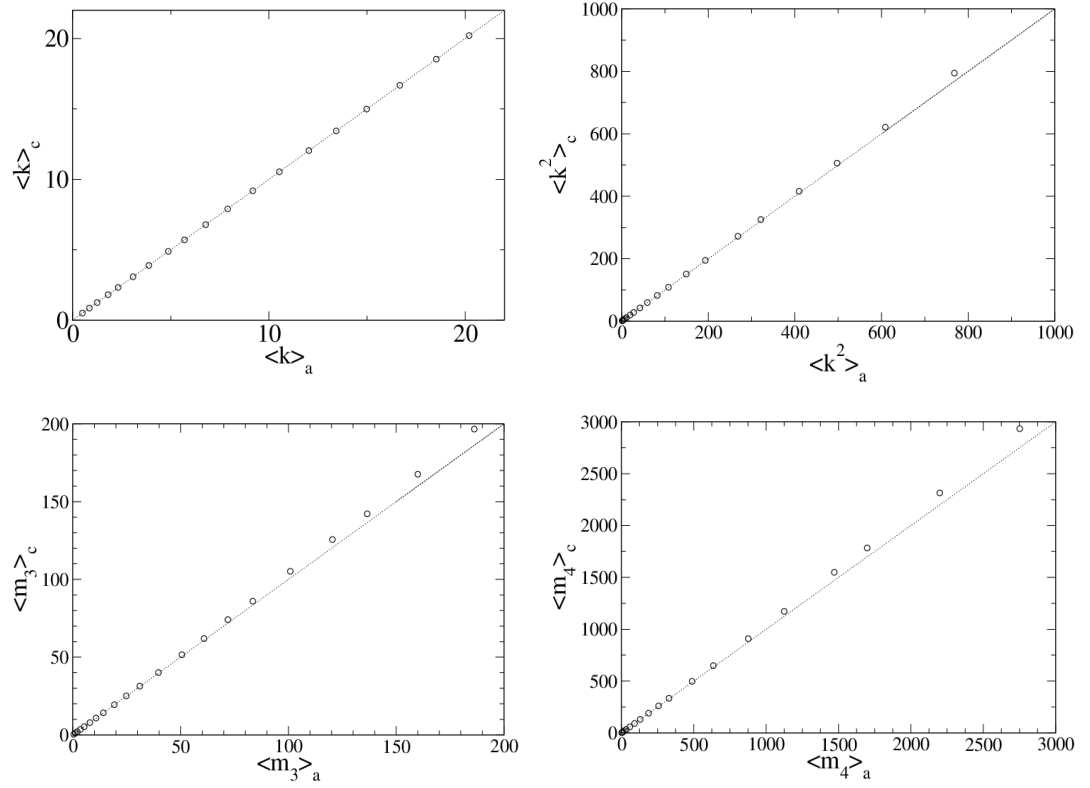




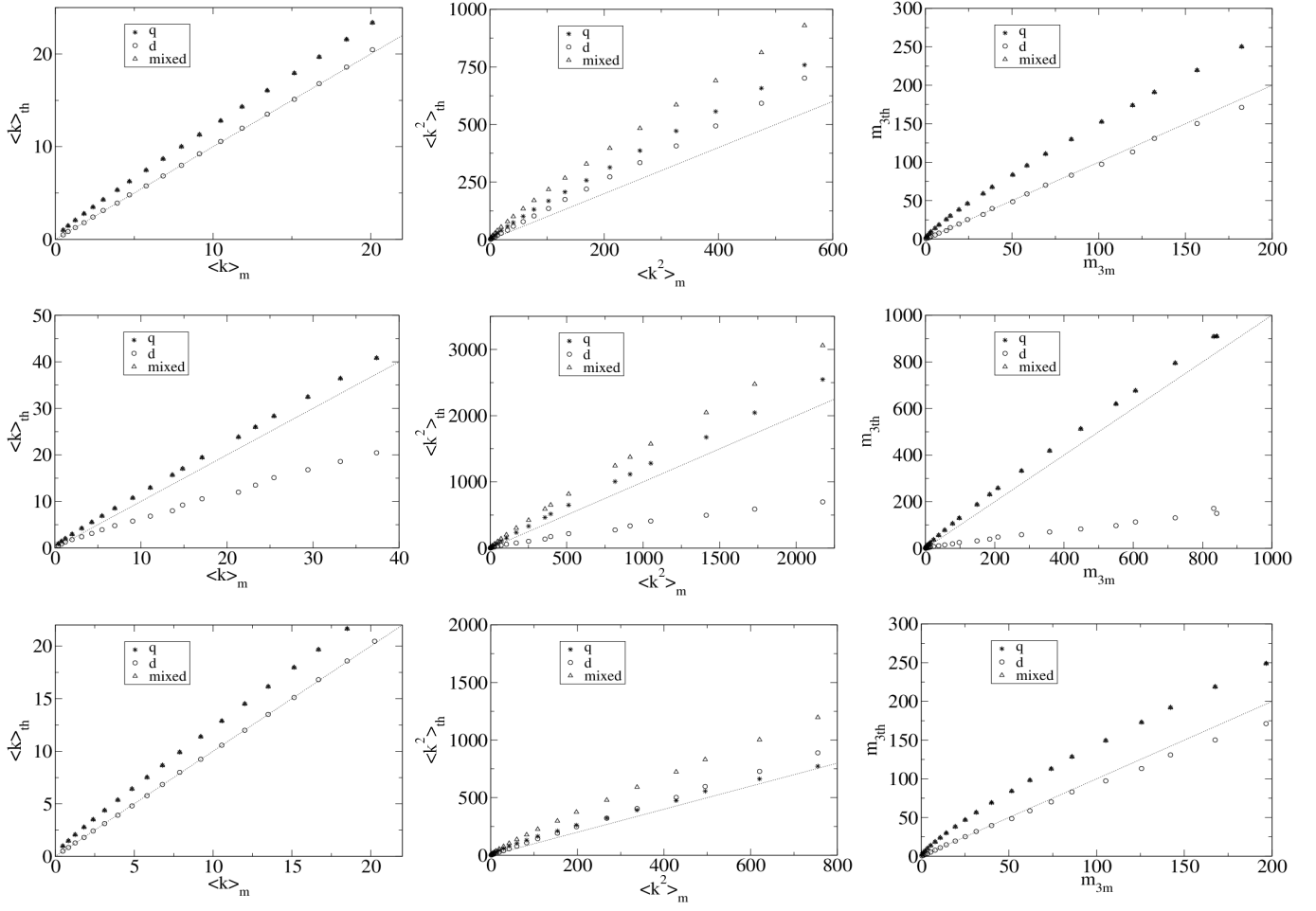
**Figure 3.** Symbols:  $\langle k \rangle$ ,  $\langle k^2 \rangle$ ,  $m_3$  and  $m_4$  as measured in synthetic graphs  $\mathbf{c}$  drawn from (11) with  $N = 3000$ , shown versus corresponding values found in the binary graphs  $\mathbf{a}$  drawn from (10). Bipartite interaction graphs  $\xi$  are drawn from (1), with complex size distributions  $P(q)$  that are Poissonian (left panels) or power law (right panels). Dotted lines: the diagonals (shown as guides to the eye). As expected, the values measured in the weighted graphs  $\mathbf{c}$  are consistently higher than in the binary ones, but one finds that these deviations get smaller for increasing network sizes  $N$ .



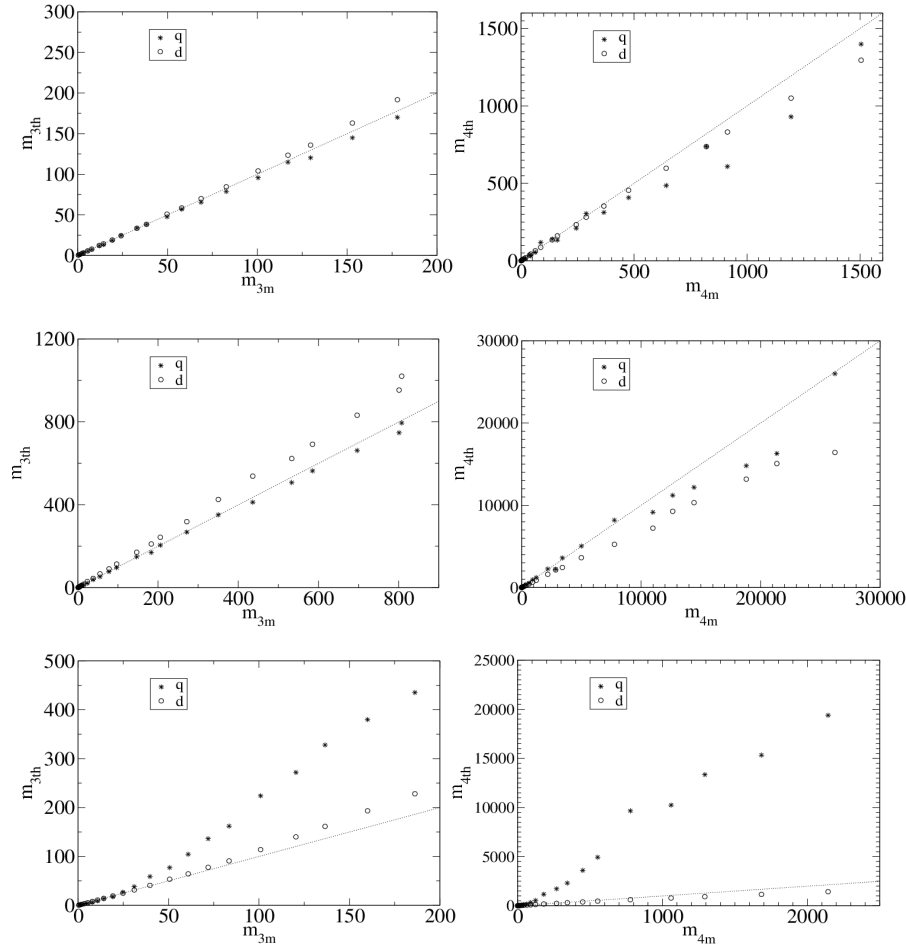
**Figure 4.** Symbols: theoretical  $\langle \dots \rangle_{th}$  versus measured  $\langle \dots \rangle_m$  values of observables  $\langle k \rangle$ ,  $\langle k^2 \rangle$ ,  $m_3$  and  $m_4$  in synthetic random graphs  $\mathcal{C}$  with  $N = 3000$ , defined via (1,11) for a power-law distributed promiscuity distribution  $P(d)$ . Theoretical values are given by formulae (63) for  $\langle k \rangle$ , (57) for  $\langle k^2 \rangle$ , (48), (59) and (66) for  $m_3$  and (49) and (66) for  $m_4$ . Dotted lines: the diagonals (shown as guides to the eye).



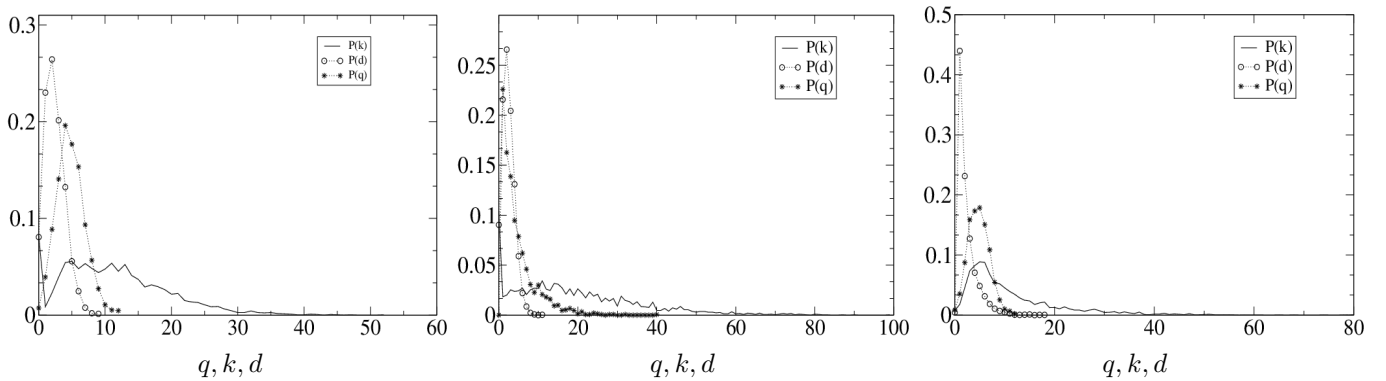
**Figure 5.** Symbols:  $\langle k \rangle$ ,  $\langle k^2 \rangle$ ,  $m_3$  and  $m_4$  as measured in synthetic graphs  $\mathbf{c}$  drawn from (11) with  $N = 3000$ , shown versus corresponding values found in the binary graphs  $\mathbf{a}$  drawn from (10). Bipartite interaction graphs  $\xi$  are drawn from (3), with protein promiscuity distributions  $P(d)$  that have a power law form. Dotted line: the diagonals (shown as guides to the eye). As expected, the values measured in the weighted graphs  $\mathbf{c}$  are consistently higher than in the binary ones, but these deviations get smaller for increasing network sizes  $N$ .



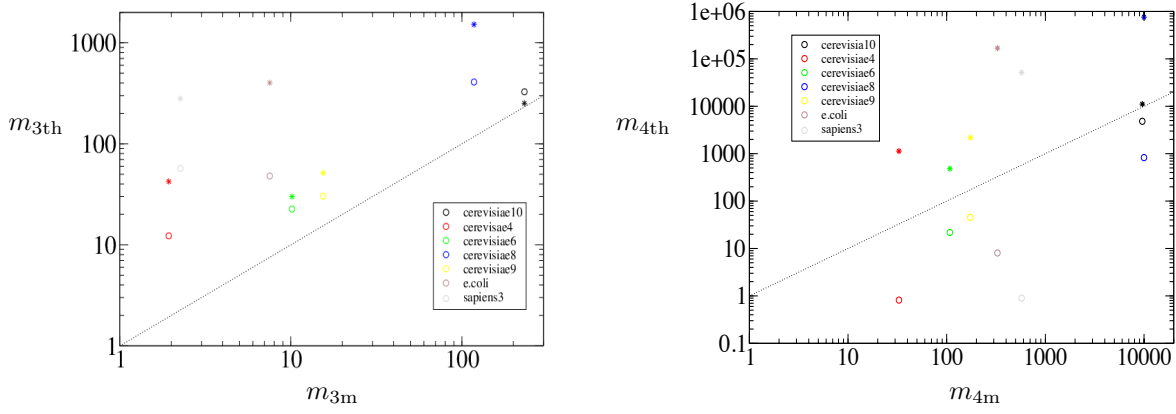
**Figure 6.** Symbols: theoretical  $\langle \dots \rangle_{th}$  versus measured  $\langle \dots \rangle_m$  values of observables  $\langle k \rangle$ ,  $\langle k^2 \rangle$ , and  $m_3$  in synthetic random graphs **a** with  $N = 3000$  and  $\alpha = 0.5$ , generated either via random wiring (top panels),  $q$ -preferential attachment (middle panels) or  $d$ -preferential attachment (bottom panels). Dotted lines: the diagonals (shown as guides to the eye).



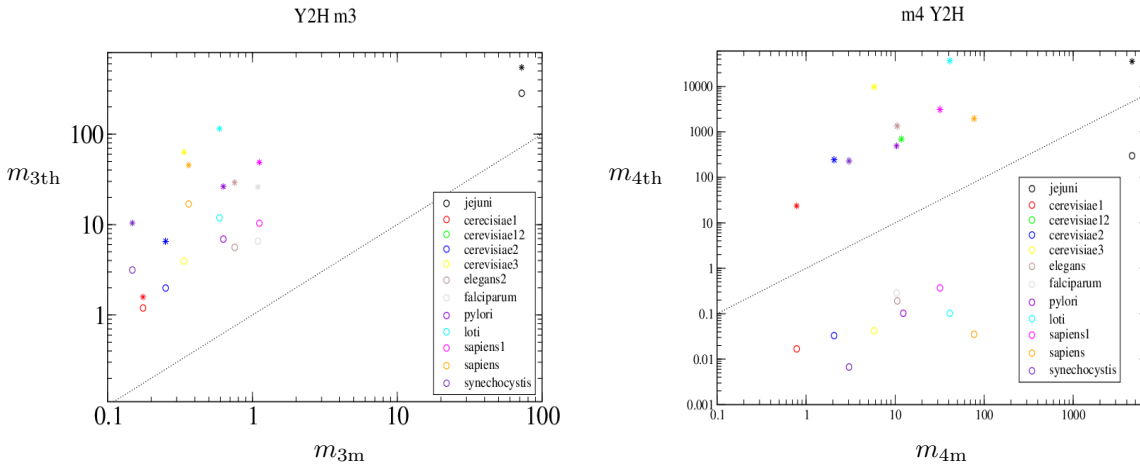
**Figure 7.** Predicted versus real  $m_3$  (left) and  $m_4$  (right) for random bi-partite graphs with  $N = 3000$  and  $\alpha = 0.5$  generated via random wiring (top panels),  $q$  preferential (middle panels) and  $d$  preferential (bottom panel), calculated by using formulae (40), (41), (66) and observables appearing in the formulae computed directly from the network.



**Figure 8.** Distributions  $P(q)$  of complex sizes,  $P(d)$  or protein promiscuities, and  $p(k)$  of the degrees in  $\mathbf{a}$  (distinguished by markers whom in the panel legends), for random bi-partite graphs with  $N = 3000$ ,  $\alpha = 0.5$  and  $\langle q \rangle = 4.8$ , which have been generated either via random wiring (left), via  $q$ -preferential attachment (middle), or via  $d$ -preferential attachment (right).



**Figure 9.** Left: theoretical predictions  $m_{3th}$  for the densities of length-3 loops in the PINs, as obtained from the  $q$ -ensemble (stars) and the  $d$ -ensemble (circles), plotted versus the values  $m_{3m}$  measured in the different MS datasets. Right: theoretical predictions  $m_{4th}$  for the densities of length-4 loops in the same PINs, obtained from the  $q$ -ensemble (stars) and the  $d$ -ensemble (circles), plotted versus the measured values  $m_{4m}$ . The diagonals are shown as guides to the eye.



**Figure 10.** Left: theoretical predictions  $m_{3th}$  for the densities of length-3 loops in the PINs, as obtained from the  $q$ -ensemble (stars) and the  $d$ -ensemble (circles), plotted versus the values  $m_{3m}$  measured in the different Y2H datasets. Right: theoretical predictions  $m_{4th}$  for the densities of length-4 loops in the same PINs, obtained from the  $q$ -ensemble (stars) and the  $d$ -ensemble (circles), plotted versus the measured values  $m_{4m}$ . The diagonals are shown as guides to the eye.